Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

# Neural Network Learning for Robust Speech Recognition

**Dissertation**
submitted to the Universität Hamburg
with the aim of achieving a doctoral degree at the Faculty of Mathematics, Informatics and Natural Sciences, Department of Informatics.

**Leyuan Qu**
Hamburg, 2021

Submission of the thesis:
18.10.2021

Day of oral defense:
15.12.2021

Dissertation Committee:

Prof. Dr. Stefan Wermter (advisor)
Dept. of Computer Science
Universität Hamburg, Germany

Prof. Dr. Timo Gerkmann (reviewer)
Dept. of Computer Science
Universität Hamburg, Germany

Prof. Dr. Jianwei Zhang (chair)
Dept. of Computer Science
Universität Hamburg, Germany

To my parents, my sisters and my girlfriend.

# Abstract

Recently, end-to-end architectures have dominated the modeling of Automatic Speech Recognition (ASR) systems. Conventional systems usually consist of independent components, like an acoustic model, a language model and a pronunciation model. In comparison, end-to-end ASR approaches aim to directly map acoustic inputs to character or word sequences, which significantly simplifies the complex training procedure. Plenty of end-to-end architectures have been proposed, for instance, Connectionist Temporal Classification (CTC), Sequence Transduction with Recurrent Neural Networks (RNN-T) and attention-based encoder-decoder, which have accomplished great success and achieved impressive performance on a variety of benchmarks or even reached human level on some tasks.

However, although advanced deep neural network architectures have been proposed, in adverse environments, the performance of ASR systems suffers from significant degradation because of environmental noise or ambient reverberation. To improve the robustness of ASR systems, in this thesis, we address the research questions and conduct experiments from the following perspectives:

Firstly, to learn more stable visual representations, we propose LipSound and LipSound2 and investigate to what extent the visual modality contains semantic information that can benefit ASR performance. The LipSound/LipSound2 model consists of an encoder-decoder with an location-aware attention architecture and directly transforms mouth or face movement sequences to low-level speech representations, i.e. mel-scale spectrograms. The model is trained in a crossmodal self-supervised fashion and does not require any human annotations since the model inputs (visual sequences) and outputs (audio signals) are naturally paired in videos. Experimental results show that the LipSound model not only generates quality mel-spectrograms but also outperforms state-of-the-art models on the GRID benchmark dataset in speaker-dependent settings. Moreover, the improved LipSound2 model further verifies the effectiveness on generalizability (speaker-independent) and transferability (Non-Chinese to Chinese) on large vocabulary continuous speech corpora.

Secondly, to exploit the fact that the image of a face contains information about the person's speech sound, we incorporate face embeddings extracted from a pretrained model for face recognition into the target speech separation model, which guide the system for predicting a target speaker mask in the time-frequency domain. The experimental results show that a pre-enrolled face image is able to benefit separating expected speech signals. Additionally, face information is complementary to voice reference. Further improvement can be achieved when com-

bining both face and voice embeddings.

Thirdly, to integrate domain knowledge, i.e. articulatory features (AFs) into end-to-end learning, we present two approaches: (a) fine-tuning networks which reuse hidden layer representations of AF extractors as input for ASR tasks; (b) progressive networks which combine articulatory knowledge by lateral connections from AF extractors. Results show that progressive networks are more effective and accomplish a lower word error rate than fine-tuning networks and other baseline models.

Finally, to enable end-to-end ASR models to acquire Out-of-Vocabulary (OOV) words, instead of just fine-tuning with the audio containing OOV words, we propose to rescale loss at sentence level or word level, which encourages models to pay more attention to unknown words. Experimental results reveal that fine-tuning the baseline ASR model with loss rescaling and L2/EWC (Elastic Weight Consolidation) regularization can significantly improve the recall rate of OOV words and efficiently overcome the model suffering catastrophic forgetting. Furthermore, loss rescaling at the word level is more stable than the sentence level method and results in less ASR performance loss on general non-OOV words and the LibriSpeech dataset.

In sum, this thesis contributes to the robustness of ASR systems by leveraging additional visual sequences, face information and domain knowledge. We achieve significant improvement on speech reconstruction, speech separation, end-to-end modeling and OOV word recognition tasks.

# Zusammenfassung

In jüngster Zeit haben End-to-End-Architekturen die Modellierung von automatischen Spracherkennungssystemen (ASR) dominiert. Konventionelle Systeme bestehen in der Regel aus unabhängigen Komponenten, wie einem akustischen Modell, einem Sprachmodell und einem Aussprachemodell. Im Vergleich dazu zielen End-to-End-ASR-Ansätze darauf ab, akustische Eingaben direkt auf Zeichen- oder Wortfolgen abzubilden, was das komplexe Trainingsverfahren erheblich vereinfacht. Es wurden zahlreiche End-to-End-Architekturen vorgeschlagen, z. B. Connectionist Temporal Classification (CTC), Sequence Transduction with Recurrent Neural Networks (RNN-T) und aufmerksamkeitsbasierte Encoder-Decoder, die große Erfolge erzielt und bei einer Vielzahl von Benchmarks beeindruckende Leistungen erbracht oder bei einigen Aufgaben sogar menschliches Niveau erreicht haben.

Obwohl fortschrittliche Architekturen für tiefe neuronale Netze vorgeschlagen wurden, leidet die Leistung von ASR-Systemen in ungünstigen Umgebungen unter Umgebungsgeräuschen oder Nachhall. Um die Robustheit von ASR-Systemen zu verbessern, gehen wir in dieser Arbeit diesen Fragen nach und führen Experimente aus den folgenden Perspektiven durch:

Erstens: um stabilere visuelle Repräsentationen zu erlernen, stellen wir LipSound und LipSound2 vor, um zu untersuchen, inwieweit die visuelle Modalität semantische Informationen enthält, die die ASR-Leistung verbessern können. Das LipSound/LipSound2-Modell besteht aus einem Encoder-Decoder mit ortsbezogener Aufmerksamkeitsarchitektur und transformiert Mund- oder Gesichtsbewegungssequenzen direkt in Low-Level-Sprachrepräsentationen, d.h. in Mel-Spektrogramme. Das Modell wird in einer crossmodalen, selbstüberwachten Art und Weise trainiert und benötigt keine menschlichen Annotationen, da die Modelleingänge (visuelle Sequenzen) und -ausgänge (Audiosignale) auf natürliche Weise in Videos gepaart sind. Das LipSound-Modell erzeugt nicht nur qualitativ hochwertige Mel-Spektrogramme, sondern übertrifft auch die modernsten Modelle auf dem GRID-Benchmark-Datensatz in sprecherabhängigen Einstellungen. Darüber hinaus bestätigt das verbesserte LipSound2-Modell die Effektivität in Bezug auf Generalisierbarkeit (sprecherunabhängig) und Übertragbarkeit (Nicht-Chinesisch auf Chinesisch) auf kontinuierliche Sprachkorpora mit großem Wortschatz.

Zweitens: Um die Tatsache auszunutzen, dass das Bild eines Gesichts Informationen über den Sprachklang der Person enthält, integrieren wir Gesichtseinbettungen, die aus einem vortrainierten Modell für die Gesichtserkennung extrahiert werden, in das Zielsprachenseparationsmodell, das das System zur Vorhersage einer Zielsprechermaske im Zeit-Frequenz-Bereich anleitet. Die experimentellen Ergeb-

nisse zeigen, dass ein vorher eingebettetes Gesichtsbild in der Lage ist, die erwarteten Sprachsignale besser zu trennen. Außerdem ist die Gesichtsinformation komplementär zur Stimmreferenz. Eine weitere Verbesserung kann durch die Kombination von Gesichts- und Stimmeinbettung erreicht werden.

Drittens stellen wir zwei Ansätze vor, um Domänenwissen, d.h. artikulatorische Merkmale (AFs), in das End-to-End-Lernen zu integrieren: (a) Feinabstimmungsnetzwerke, die Repräsentationen der versteckten Schicht von AF-Extraktoren als Input für ASR-Aufgaben wiederverwenden; (b) progressive Netzwerke, die artikulatorisches Wissen durch laterale Verbindungen von AF-Extraktoren kombinieren. Die Ergebnisse zeigen, dass progressive Netzwerke effektiver sind und eine niedrigere Wortfehlerrate erreichen als Feinabstimmungsnetzwerke und andere Basismodelle.

Um schließlich ASR-Modelle in die Lage zu versetzen, Wörter außerhalb des Vokabulars (OOV) zu erfassen, anstatt nur eine Feinabstimmung mit den Audiodaten vorzunehmen, die OOV-Wörter enthalten, schlagen wir vor, die Verlustfunktion auf Satz- oder Wortebene neu zu skalieren, was die Modelle dazu anregt, unbekannten Wörtern mehr Aufmerksamkeit zu schenken. Die experimentellen Ergebnisse zeigen, dass die Feinabstimmung des ASR-Basismodells mit Verlust-Reskalierung und L2/EWC-Regularisierung die Wiedererkennungsrate von OOV-Wörtern erheblich verbessern und beim Modell katastrophales Vergessen reduzieren kann. Darüber hinaus ist die Verlust-Reskalierung auf Wortebene stabiler als auf Satzebene und führt zu geringeren ASR-Leistungsverlusten bei allgemeinen Nicht-OOV-Wörtern und bereits gelernten Aufgaben.

Zusammenfassend trägt diese Arbeit zur Robustheit von ASR-Systemen bei, indem zusätzliche visuelle Sequenzen, Gesichtsinformationen und Domänenwissen genutzt werden. Wir erreichen signifikante Verbesserungen bei der Sprachrekonstruktion, der Sprachseparation, der End-to-End-Modellierung und der OOV-Worterkennung.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

In the last decades, automatic speech recognition (ASR) has received more and more attention, since speech is perceived as the most natural and ideal modality for human-machine interaction. Typically, the conventional ASR system consists of an acoustic model that maps acoustic inputs to phoneme sequences, and a language model that transforms the candidates' outputs from the acoustic model into grammatically and semantically correct words. However, the two parts are built separately.

Conventional acoustic models learn the information of temporal and semantic variability in speech signals with hidden Markov models (HMMs) and use Gaussian mixture models (GMMs) to map a window of acoustic features, such as Mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive coefficients (PLPs) [75], to HMM states. With the advance of machine learning algorithms and computational hardware, replacing GMM with DNN has been a trend since deep neural networks (DNNs) can learn more effective representations from audio waveforms and acquire longer context information that helps to disambiguate the errors caused by local biases. The DNN-HMM hybrid model significantly promotes the development of ASR and makes it possible to be applied in the real world. Although the DNN-HMM method has been explored extensively, it still has many inherent limitations and requires frame level labels that have to be obtained from a GMM-HMM model iteratively. Besides, training a good ASR model is complex due to the requirement of domain knowledge and multi-step optimization.

To simplify this complex paradigm, end-to-end learning approaches, for example connectionist temporal classification (CTC) [61], RNN-Transducer [59] and

attention-based encoder-decoder [24], have been proposed, which replace hand-designed feature engineering and jointly learn all components in a single architecture (see detailed review in Section 3.2). These approaches are transformed into computational flow graphs that can be optimized by backpropagation in a simple end-to-end training process. Besides, end-to-end models can naturally handle sequences of arbitrary lengths and directly optimize the word error rate. End-to-end models have made substantial progress recently on a variety of benchmarks [6, 34, 14, 24] or reached human parity on some tasks [206].

Although advanced deep neural network architectures have been proposed, in realistic or adverse environments, the performance of ASR systems suffers from significant degradation because of environmental noise [88, 15, 70, 200]. Plenty of methods have been proposed to suppress noise from software algorithms, for example, speech enhancement [142, 139] and speech separation [192, 116], and hardware devices, for instance, multi-channel microphone array [166, 196]. However, the most existing methods for improving ASR robustness are fully data-driven, which is hard to integrate domain knowledge. Besides, the hardware-dependent approaches are expensive and difficult to generalize to mobile devices. In this thesis, we explore how to leverage additional visual information (mouth movement sequences and face images) and domain knowledge (articulatory features) to benefit ASR systems.

## 1.2 Research Objectives

Inspired by human bimodal perception [18] in which both sight and sound are used to improve the comprehension of speech, a lot of effort has been spent to improve ASR model robustness by leveraging visual information, for example, integrating simultaneous lip movement sequences into speech recognition [36, 1], separating target speech signals with a static face image for speech separation [155, 38] and grounding speech recognition with visual objects and scene information [125, 66]. Multi-modal audio-visual methods achieve significant improvement over single modality models in adverse environments since the visual signals are invariant to acoustic noise and complementary to audio representations [118]. Moreover, the visual contribution becomes more important as the acoustic signal-to-noise ratio is decreased [170].

However, many experimental results reveal that limited or minor improvement is obtained when incorporating visual information into speech recognition systems, for example, only 4.7% relative improvement on word error rate when combining additional video streams (20.5% vs 21.5%) [53]. Additionally, visual speech recog-

nition (or lip reading) that only uses visual sequences as inputs still cannot be competitive with audio speech recognition (48.5% vs 21.5%) [53]. Phoneticians found that the visual modality carries less relevant information for recognition than audio [54]. Furthermore, some phonemes are visually identical but different and discriminative in audio [119]. For example, in English, the minimal pairs /b/ and /p/, /b/ is a voiced sound while /p/ is an unvoiced sound. They are different in audio-based speech representation, while the two phonemes are modeled as the same unit in traditional lip reading systems, since /b/ and /p/ are produced with the same visually apparent lip and tongue movement.

Whether the visual similarity is distinguishable for the speech processing community is still an open question and limited research has been done quantitatively, which leads to the first research question in this thesis:

**Q1: How can neural networks reconstruct intelligible speech from mouth or face movement sequences when speech signals are noisy or not available?**

To answer this question, an attention-based encoder-decoder architecture is proposed to directly map videos to corresponding audio to maximize the semantic information in visual streams. Experiments are conducted in Chapter 4 and 5.

Current robust ASR systems usually require a speech separation model that transforms noisy speech into clean signals or representations to improve ASR performance. Conventional target speaker separation systems condition on a voice reference to recover a clean speech signal from a speech mixture, however, the auxiliary voice information has to be pre-enrolled, which is hard to meet in most real-word scenarios.

Inspired by the finding by neuroscientists [16, 122] and psychologists [23, 165] that there is a strong relationship between faces and voices and sometimes humans can even infer what one's voice sounds like by only seeing the face, or vice versa, researchers in computer science have conducted a large number of studies on learning face and voice association, for example, reconstructing human faces by only conditioning on speech signals [136] or learning joint or sharing face-voice embedding space for tasks of crossmodal biometric retrieval or matching [134]. In comparison, a face image is easy to obtain by pre-enrollment or capture with a camera in real-time. We would thereby propose the second research question,

**Q2: How can we improve the speech separation/recognition performance with a face image?**

To explore this question, we incorporate face embeddings extracted from a pre-trained model for face recognition into the target speech separation task, which guides the system in predicting a target speaker mask in the time-frequency domain. Experiments are presented in Chapter 6.

Although end-to-end architectures dramatically simplify the training procedure of ASR systems, this "black box" is fully in a data-driven fashion where the intrinsic mechanism is inexplicable and uncontrollable, which leads to many concerns, for example, the system is easily fooled or attacked by adversarial inputs [86]. It is challenging to integrate domain knowledge into these models. Consequently, the third question explored in this thesis is:

**Q3: How can we incorporate domain knowledge into end-to-end architectures to improve the robustness of ASR systems?**

Articulatory features (AFs), also known as distinctive phonetic features, are used to represent the movement of different articulators, such as lips and tongue, during speech production, which can improve the performance of ASR systems by systematically accounting for coarticulation, speaking styles and other variabilities, especially in a noisy scenario [94]. To increase the robustness of ASR systems, in this thesis, we exploit several manners to integrate the domain knowledge, i.e. AFs, into end-to-end architectures in Chapter 7.

Additionally, the end-to-end models rely heavily on data and perform poorly on words out-of-vocabulary (OOV) or rarely existing in training data, for example, trending words and new named entities. Since it is expensive to collect labeled OOV speech data for ASR model training, current approaches for solving OOV problems mainly focus on language model (ML) or post-processing, for instance, user-dependent language model [21, 121], LM rescoring [65] and finite-state transducer lattice extension [219]. However, the post-processing techniques only obtain limited improvement and it is hard to tackle the root causes acoustically. The last question we investigate in Chapter 8 is:

**Q4: How can we keep ASR models robust when adding OOV words?**

## 1.3 Novelty and Contribution

The main novelties and contributions of this thesis are as follows:

1. We propose novel encoder-decoder LipSound and LipSound2 architectures that directly map mouth or face movement sequences to mel-scale spectro-

grams for speech reconstruction, which does not require any human annotations. Previous work only focuses on speaker-dependent settings where the speakers in training sets are used for testing as well. Our proposed method can not only significantly outperform the state-of-the-art models in speaker-dependent settings but also can successfully generalize to any unseen speakers not existing in the training procedure (speaker-independent). We make significant progress towards the real-world application of speech reconstruction.

2. Compared to previous work that only conducts experiments on small vocabulary or artificial grammar datasets, we innovatively propose the approach of cross-modal self-supervised pre-training for speech reconstruction. The proposed LipSound2 model is firstly pre-trained on the large-scale multilingual VoxCeleb2 dataset, then the pre-trained model is fine-tuned on the domain-specific datasets. In this way, we successfully generalize our model to spontaneous speech and large-scale vocabulary cases, which puts the speech reconstruction technique one step closer to the realistic scenarios.

3. We propose a novel approach by integrating pre-enrolled face information into the target speech separation task. Our model avoids the speaker permutation problem and the problem of an unknown number of source speakers, which audio-only approaches suffer from. In addition, different from the conventional audio-visual speech separation methods which heavily rely on the temporal information from the visual sequences, our system can also be easily adapted to those devices without cameras or to scenarios where no simultaneous visual streams are available.

4. We present two approaches to combine domain knowledge, i.e. AFs, into end-to-end learning. First, fine-tuning neural networks are proposed to concatenate hidden layer outputs of AF extractors as inputs to another RNN for ASR. Second, a progressive neural network with lateral connections from AF extractors is proposed to integrate articulatory knowledge into an end-to-end architecture.

5. We investigate using synthetic speech to boost the ASR model on the recognition of OOV words, for instance, trending words or new named entities. Instead of just fine-tuning with audio containing OOV words, we propose to rescale loss at sentence level or word level, which encourages models to pay more attention to unknown words.

## 1.4 Thesis Organization

This thesis is structured into 9 chapters. The methodology and related work are presented in Chapter 2 and Chapter 3 respectively.

- To exploit the benefit of visual modality for speech recognition:
  - In Chapter 4, we propose the LipSound architecture and give the details of the model front-end and back-end, followed by experiments on speech reconstruction and lip reading on GRID dataset. The comparison of model performance with previous work is presented as well in the discussion.
  - In Chapter 5, we propose LipSound2 to further explore to what extent the large-scale crossmodal self-supervised pre-training can benefit speech reconstruction in generalizability (speak-independent) and transferability (Non-Chinese to Chinese). In addition, related work on lip to speech reconstruction and lip reading is reviewed in this chapter.
- To learn voice identity information from a face image:
  - Chapter 6 begins with the development review of target speech separation and learning associations between faces and voices in recent years. Then, the proposed face-guided target speech separation architecture is demonstrated. Model performance is evaluated on speech separation and speech recognition experiments.
- To incorporate domain knowledge into end-to-end ASR architectures:
  - Chapter 7 introduces the extraction of articulatory features and 2 ways to integrate them into end-to-end learning.
- To enable ASR models to acquire OOV words:
  - In Chapter 8, related work is firstly reviewed on the recognition of OOV words in end-to-end ASR models and data augmentation with synthetic audio for ASR. The proposed approaches of loss rescaling on word level and sentence level are then demonstrated.

Finally, we conclude this thesis and give the answers to our research questions in Chapter 9.

# Chapter 2

# Methodology of Neural Networks and End-to-End ASR

## 2.1 Introduction

In this chapter, we describe some foundations of conventional neural networks (CNNs) and Dilated Convolutional Neural Networks, which are usually used in the first few layers of our proposed models since they are good at capturing the local variation from images or audio. Then Long Short-term Memory Networks and their variant Gated Recurrent Units (GRUs) are introduced, which normally follow CNNs to learn long-distance dependence. We also demonstrate the connectionist temporal classification (CTC) loss function that is used in the end-to-end ASR systems, i.e. Jasper [110] in Chapter 5, 6 and DeepSpeech2 [6] in Chapter 4, 7. Finally, location-aware attention is presented, which is utilized to form a bridge between the encoder and the decoder in our proposed LipSound and LipSound2 architectures.

## 2.2 Convolutional Neural Networks

Different from feed-forward neural networks, where each node connects via different weights and has a different bias, convolutional neural networks (CNNs) [106] capture local variations by leveraging matrix multiplication, which is utilized specifically in computer vision tasks, for example, image classification and semantic segmentation. The CNN computational mechanism is shown in Figure 2.1, where the filter is 2D size and slides over the input matrix orderly.

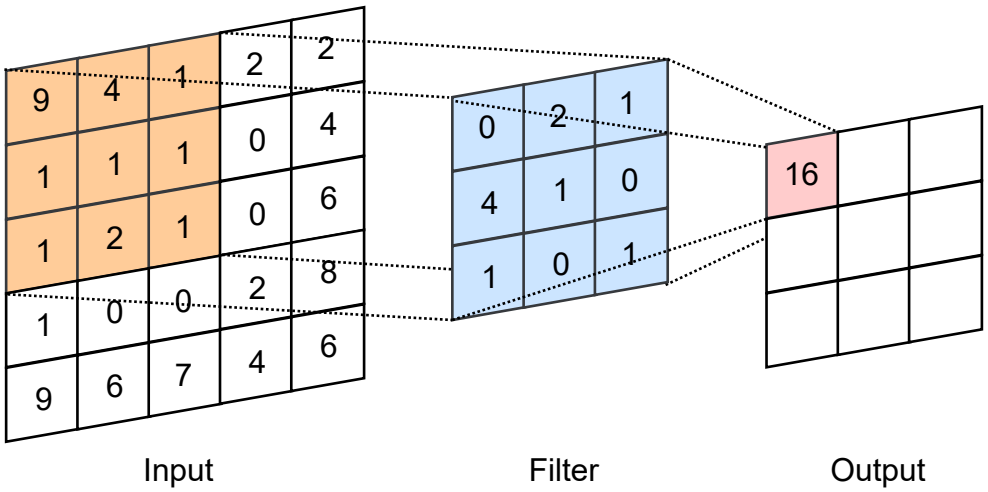


Figure 2.1: The computational mechanism of CNN with 2D filter size.

The CNNs are usually used with other techniques, like padding the boundary of input matrices with zero or pooling outputs with max/mean values for downsampling. Different sizes of filters are used for different input vectors, for instance, 1D filter for text sequences, 2D filters for images and 3D filters for videos. We develop 2D (frequency and time domains) CNNs on the input Mel-scale spectrograms at the first two layers of DeepSpeech2 [6] in Chapter 4 and 7, since the convolution layers not only reduce temporal variability in the time domain but also normalize speaker variance in the frequency domain. Besides, 1D convolution [110] is also used to process 2D Mel-spectrograms in Chapter 5 and 6, where filters move in the time domain and the number of filters equals the frequency bands of input speech features.

## 2.3 Dilated Convolutional Neural Networks

Regular CNNs are good at capturing local variation but perform poor on long-distance context dependence. It can only be solved by increasing filter size, for example, from a 3×3 to a 5×5 filter. However, the increase in filter size leads to model parameter explosion and makes it difficult to converge. Alternatively, dilated convolutional neural networks are thereby proposed to increase model receptive fields with pixel skipping and keep the same amount of parameters as regular

CNNs. As shown in Figure 2.2, the dilated CNN layer takes every other value from the input matrix to multiply with the filter when using a dilation factor of 1, which can efficiently enable the model to see more data and longer context information.

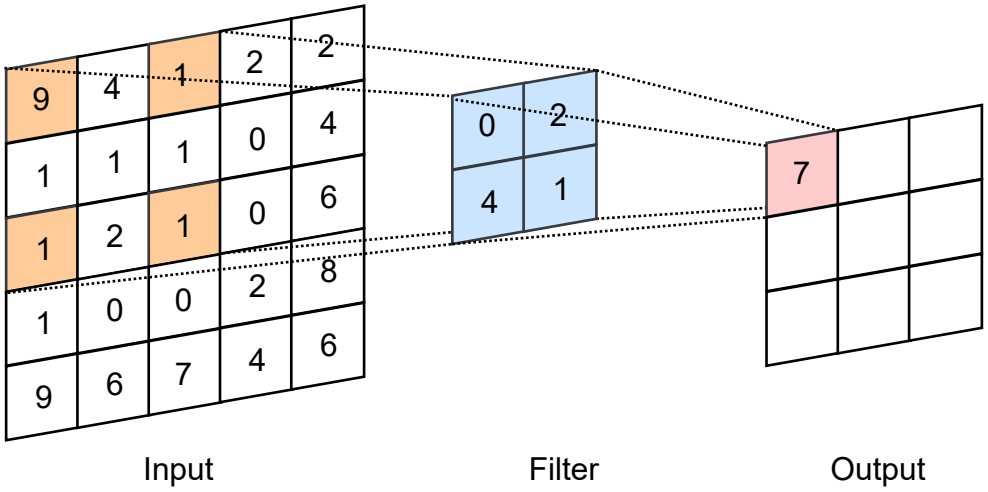


Figure 2.2: The computational mechanism of a dilated CNN with 2D filter size and dilation factor of 1.

## 2.4 Long Short-term Memory Networks

To tackle sequence problems, for instance, machine translation and speech recognition, long short-term memory networks [77] as a type of recurrent neural network (RNN) have been proposed to predict future information by conditioning on previous memory. Different from the standard RNN that utilizes only one simple activation function, e.g. tanh, per cell, the LSTM cell consists of a more complex structure and several gates, i.e. an input gate, an output gate, and a forget gate. The internal structure of LSTM block is shown in Figure 2.3.

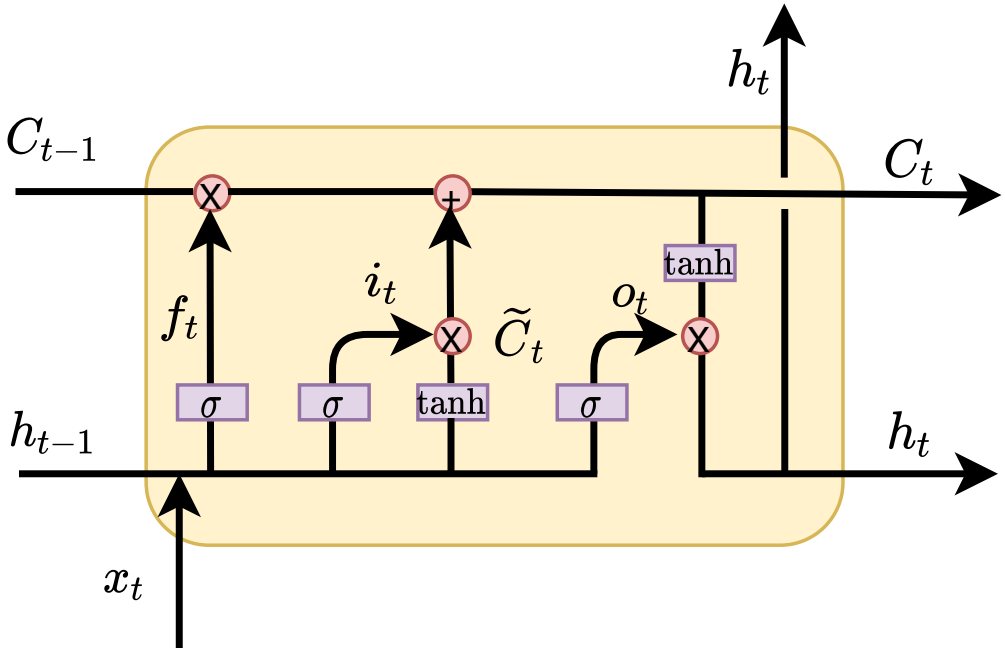


Figure 2.3: The internal structure of LSTM.

We demonstrate these gates in the following formulas.

**The input gate $i_t$:**

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.1}$$

where $h_{t-1}$ is the short term memory from last step. $x_t$ and $b_i$ are the input vector at time step $t$ and bias for matrix $W_i$ respectively.

**The forget gate $f_t$:**

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.2}$$

**The output gate $o_i$:**

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.3}$$

**The long term memory $C_t$:**

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t \tag{2.4}$$

where $\widetilde{C}_t$ is

$$\widetilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2.5}$$

**The short term memory:**

$$h_t = o_t \cdot tanh(C_t) \tag{2.6}$$

Compared to naive RNN, LSTM can alleviate the problem of gradient vanishing and gradient exploding by optionally keeping or forgetting information with the three gates.

## 2.5 Gated Recurrent Units

However, the complex structure and parameters in LSTM lead the network to gradient explosion in the training procedure sometimes. Gated recurrent units (GRUs) [32] are proposed to simplify LSTM and accelerate the training process. GRUs consist of reset gate and update gate to control the flow of information, as shown in Figure 2.4.

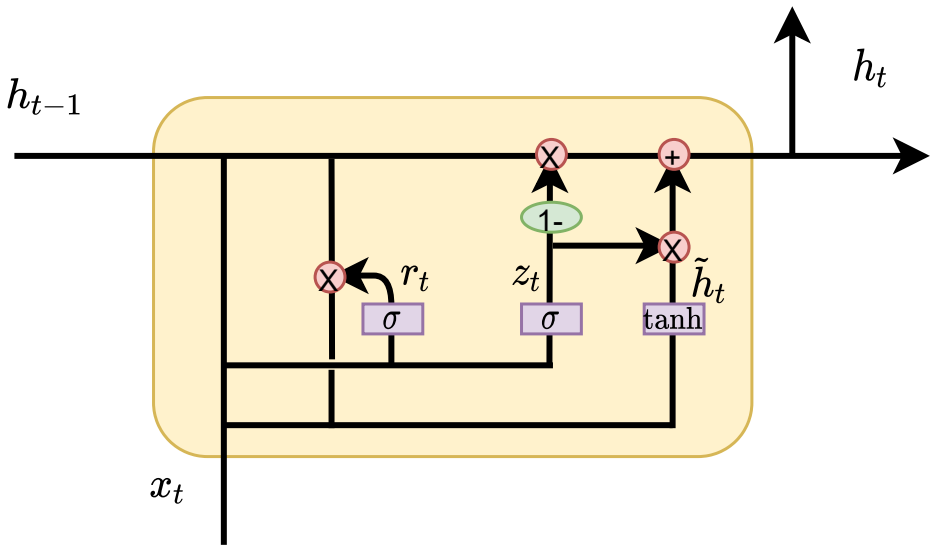


Figure 2.4: Flow diagram of a GRU cell.

**The update gate $z_t$:**

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{2.7}$$

where $h_{t-1}$ is the hidden output from last step and $x_t$ is the input vector at step $t$.

**The reset gate $r_t$:**

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{2.8}$$

**The hidden output $h_t$**

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h}_t \tag{2.9}$$

where

$$\widetilde{h}_t = tanh(W \cdot [r_t * h_{t-1}, x_t]) \tag{2.10}$$

Compared with LSTMs in which 3 gates are used, GRUs only utilize 2 gates. It has been shown that GRU cells achieve comparable performance to LSTMs but GRU cells are faster and easier to train [83].

## 2.6 Conventional ASR Models

When training ASR systems with the conventional method, frame-level labels are required, which are usually estimated by dynamic programming with GMM-HMM models building on phoneme or tri-phone units. The conventional ASR model generates the same length of token sequence as acoustic inputs, as shown in Figure 2.5, followed by a pronunciation dictionary to map the intermediate units to characters or words by leveraging a language model that is trained separately.

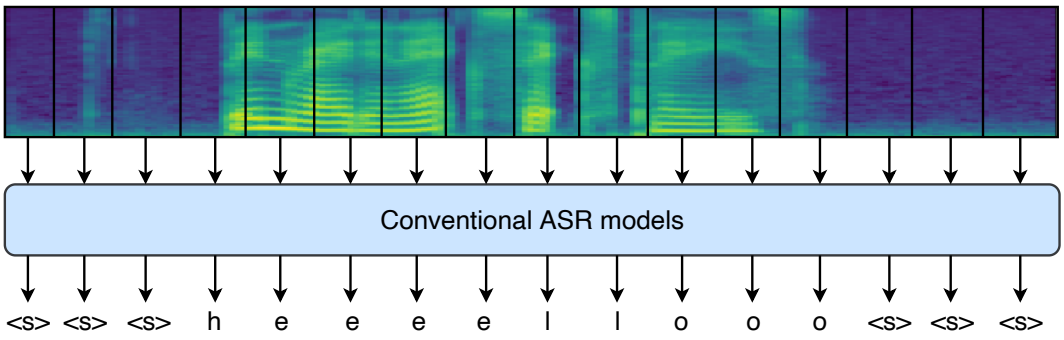


Figure 2.5: The conventional ASR system predicts one label for every input frame.

Given an input acoustic vector $\mathbf{x} = (x_0, \cdots, x_T)$, and the corresponding target label sequence $\mathbf{y} = (y_0, \cdots, y_U)$, the ASR model aims to predict an output sequence $\mathbf{y}$ with highest probability by conditioning on the input $\mathbf{x}$. The conventional ASR model aims to predict an output sequence $\mathbf{y}$ with highest probability by conditioning on the input $\mathbf{x}$, as shown in Eq. (2.11). Eq (2.12) can be obtained by utilizing Bayes' theorem where $P(X)$ is a constant and can be neglected. Thereby, the ASR system is simplified to maximize the product of $P(X|Y)$ and $P(Y)$, where $P(X|Y)$ and $P(Y)$ is called acoustic model and language model respectively.

$$
\begin{aligned}
Y^* &= argmax_Y P(Y|X) && (2.11)\\
&= argmax_Y \frac{P(X|Y) \cdot P(Y)}{P(X)} && (2.12)\\
&\propto argmax_Y \underbrace{P(X|Y)}_{\text{Acoustic Model}} \cdot \underbrace{P(Y)}_{\text{Language Model}} && (2.13)
\end{aligned}
$$

## 2.7 Connectionist Temporal Classification

Different from conventional ASR models that require precise phoneme boundary information, the connectionist temporal classification (CTC) [61] loss function inserts one blank token $\phi$ to blur the boundary. As shown in Figure 2.6, the CTC-based ASR model generates all possible candidate paths in which the blank token $\phi$ is used to segment neighboring labels. Then a carefully designed dynamic programming algorithm is used to search optimal paths and convert the frame-level token sequences to meaningful utterance by removing blank tokens and merging repeated labels.

## 2.8 Location-aware Attention

We use location-aware attention [34] to join the encoder and the decoder in our proposed LipSound models.

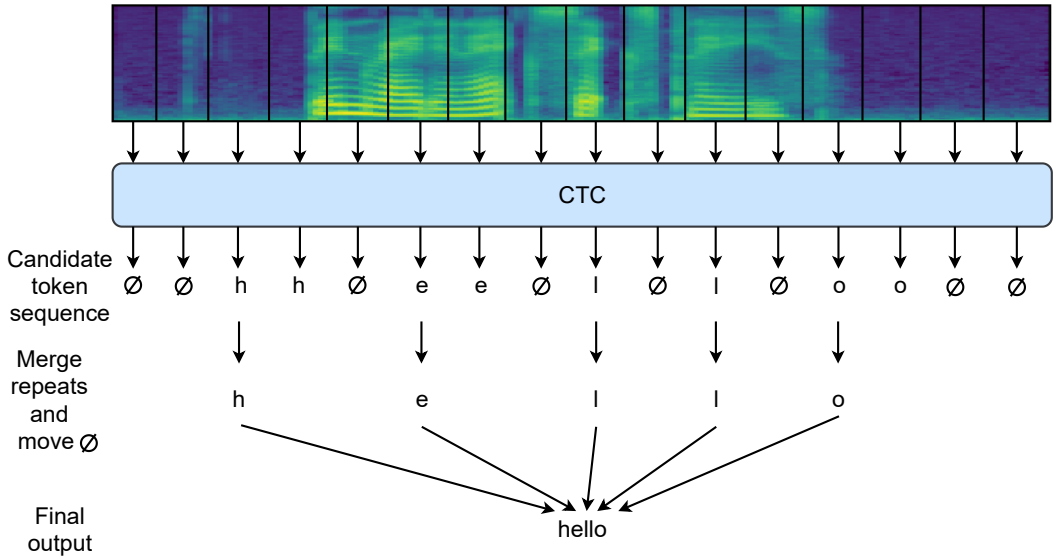


Figure 2.6: The mechanism of CTC function.

The image sequence input $i = (i_0, ..., i_n)$ is firstly embedded into the latent space representation vector $h = (h_1, ..., h_n)$ by the encoder with the same dimension $n$ in time, then the intermediate vector $h$ is decoded into the Mel-spectrogram $o = (o_0, ..., o_m)$. At time step $t$ $(0 \leq t \leq m)$, the attention weight $a_t$ can be obtained by the following equations:

$$a_t = Softmax(W \cdot tanh(M \cdot h + Q \cdot x + L \cdot y)) \tag{2.14}$$

$$x = LSTM(h \cdot a_{t-1}, p_{prenet}) \tag{2.15}$$

$$y = Conv(a_{t-1}, \sum_{0 \leq i \leq t-1} a_i) \tag{2.16}$$

$$v_t = a_t \cdot h \tag{2.17}$$

where $W, M, Q, L$ are the matrices learned by the layer of Weight FC (Fully Connected), Memory FC, Query FC and Location FC, respectively (please see

more details in Chapter 5). In Eq. (2.16), the sum of attention weights of all previous steps is integrated, which enables the current step attention to be aware of the global location and move forward monotonically. Fig. 2.7 visualizes the computational flow of the attention mechanism. The attention content vector $v_t$ can be obtained by multiplying the encoder output by the normalized attention weights (Eq. (2.17)).

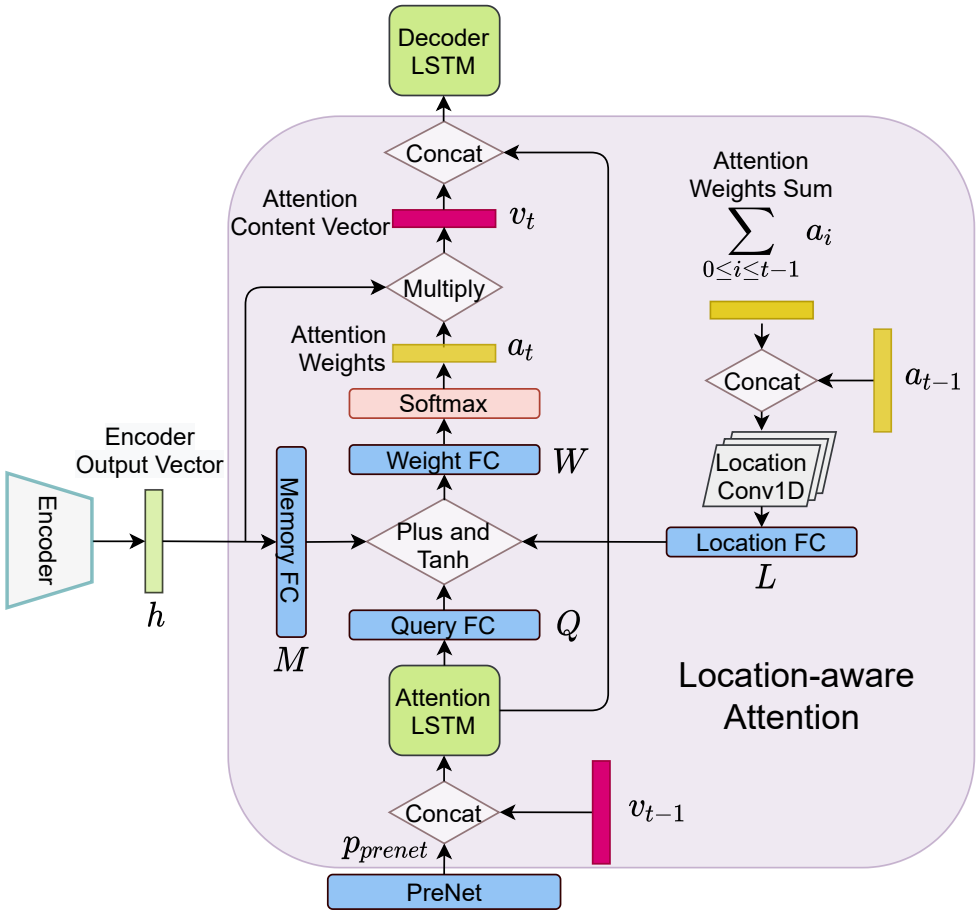


Figure 2.7: The computational flow of location-aware attention at time step $t$.

## 2.9 Summary

In this chapter, we presented some foundations that our thesis and proposed approaches are based on. We mainly introduce CNN, Dilated CNN, LSTM and GRU networks. A brief introduction about CTC loss function is provided, which enables ASR systems to be trained in an end-to-end fashion. Finally, the location-aware attention is presented, which plays an important role in encoder-decoder architecture .

# Chapter 3

# Related Work on Speech Recognition and Multi-modal Speech Processing

## 3.1 Introduction

We review related work and state-of-the-art approaches in this chapter. We firstly review end-to-end ASR models in three directions, i.e. CTC, RNN-T and attention-based encoder-decoder architecture, since the thesis aims to improve the robustness of end-to-end approaches, where the ASR systems used in Chapter 4, Chapter 5, Chapter 6 and Chapter 7 are CTC-based models and we use the hybrid CTC/attention model in Chapter 8.

Secondly, we present previous state-of-the-art approaches on lip to speech reconstruction, lip reading and self-supervised learning that are the main focuses of Chapter 4 and 5, followed by the review of the latest progress on target speech separation and learning associations between faces and voices, which are the background of Chapter 6 where we explore face-guided target speech separation systems.

Thirdly, we review related work for Chapter 7 on integrating domain knowledge into ASR systems. In this chapter, we only consider studies that involve linguistic and phonetic knowledge since this is the main focus of Chapter 7.

Then, the related work for Chapter 8 on the recognition of OOV words in end-to-end ASR models is reviewed in Section 3.9, in which we demonstrate the current solutions used in industry and academia. Besides, we also review the data augmentation method with synthetic audio for ASR modeling in Section 3.10.

## 3.2 End-to-End ASR Architectures

End-to-end ASR architectures aim to directly map acoustic observations to transcripts, such as characters and words. Different from conventional systems which separately train the acoustic model and LM with different criteria and corpora, end-to-end ASR integrates all modules into one neural network. Consequently, the training progress is significantly simplified and does not require too much domain-specific knowledge, which attracts more and more researchers from other areas, like computer vision and natural language processing, and heavily promotes the development of speech-related multi-modal and cross-model learning. There are three main categories in end-to-end learning, CTC, RNN-Transducer (RNN-T), and encoder-decoder with attention architectures. They are mainly different from the way of aligning input acoustic features and output label sequences, in other words, how to calculate the loss function between two sequences with variable length.

### 3.2.1 CTC-based End-to-End Models

The basic idea of CTC-based approaches is to bring in a special token, 'blank', which is dynamically filled in the places between modeling units. Consequently, the exact boundary information required by conventional methods is not needed anymore. Then a carefully designed dynamic programming algorithm is used to search optimal paths and convert the frame level token sequences to meaningful utterance by removing blank tokens and merging repeated labels. The CTC-based architecture attracts a lot of researchers from both industry and academia.

The Figure 3.1 shows the schematic of CTC, where the encoder generates one posterior $P(y_t|x_t)$ probability for each input acoustic frame $x_t$. Then the CTC function is used to map to the text sequence and calculate losses.

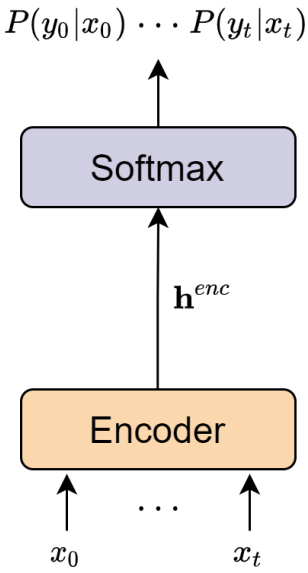


Figure 3.1: Schematic of CTC. The figure adapted from [149].

Graves et al. [61] firstly adopt the ground-breaking CTC approach in speech recognition tasks to overcome the problems, i.e. the requirement of frame level segmentation and the mapping from model outputs to ground truth labels, which mainly constrain the development of ASR systems. The novel CTC algorithm provides the possibility for breaking limitations in conventional ASR modeling. Sak et al. [163] compare the CTC modeling technique with the conventional hybrid DNN-HMM method. Besides, modeling units and context effectiveness are investigated on HMM states v.s. phones and unidirectional/bidirectional LSTMs respectively. Some training strategies, like state-level minimum Bayes risk (sMBR) and finite-state transducer (FST), help the CTC-based model achieving competitive results with DNN-HMM models. Hannun et al. [69] simplify the CTC building process by replacing LSTM with a bi-directional recurrent deep neural network (BRDNN). Moreover, a modified prefix-search decoding algorithm is proposed to completely discard the cumbersome decoding strategies used in HMM-based systems. Amodei et al. [6] conduct a comprehensive evaluation on model architecture, system optimization, and model deployment. Data amount and parameter amounts, such as layer depth and layer width, are deeply analyzed. Besides, model level [40] and data level [44] parallel training, SortaGrad and batch normalization are utilized to heavily speed up the training process. Furthermore, plenty of strategies for practical model deployment are investigated. The proposed DeepSpeech2 system not only significantly promotes the development of a CTC-based system in

academia but also dramatically boosts the landing of products in the industry. Audhkhasi et al. [11] directly build ASR models at word level units instead of on phone- or subword level which requires an external language model. The results on Switchboard and CallHome benchmarks suggest that systems training at word level are more data-hungry and easily suffer from long-tail issues. Billa et al. [19] improve the recurrent model performance with dropout regularization techniques in the LSTM-CTC structure. Zhang et al. [218] investigate the modeling units, like context-independent initial/finals and context-dependent initial/finals, for Chinese on a DFSMN-CTC-sMBR acoustic model. The results show the best performance when modeling on hybrid character-syllable units.

Besides, substituting RNN with CNN has been a growing trend since pure CNN-based architecture can dramatically reduce the training time and inference latency, which is critical for speech recognition tasks running in real-time. Jasper [110] achieves state-of-the-art results on LibriSpeech dataset by stacking the CNN-only block which consists of 1DCNN, batch normalization, ReLU, and dropout. Inspired by Jasper, Krimany et al. [101] propose QuartzNet which uses a similar fully convolutional architecture as Jasper but with significantly fewer parameters. QuartzNet enables the CTC-based end-to-end ASR model to run on mobile devices locally.

However, some shortcomings made by CTC loss function limit its application on other tasks. Firstly, the length of target labels must be shorter than input sequences to insert blank labels into target label sequences. This assumption makes it not suitable for tasks like TTS, where the outputs are larger than inputs. Secondly, the projection between inputs and outputs must be monotonic. Such limitation leads to the non-applicability of tasks like machine translation where a future word in input sequence could be aligned to an earlier word in output labels. Lastly, to simplify the calculation of path probability, CTC assumes that every output node is conditionally independent of other outputs, which makes it difficult to model the dependency between adjacent frames.

### 3.2.2 RNN-T End-to-End Models

To overcome the unreasonable assumptions of conditionally independent in CTC, RNN-T [59] was proposed. As shown in Figure 3.2, different from CTC which only focuses on acoustic sequence modeling, RNN-T uses a separate module, called predictor, to model the context information from last step $y_{u-1}$, which can be treated as an intrinsic language model, followed by a jointer to classify the concatenation of encoder and predictor outputs.

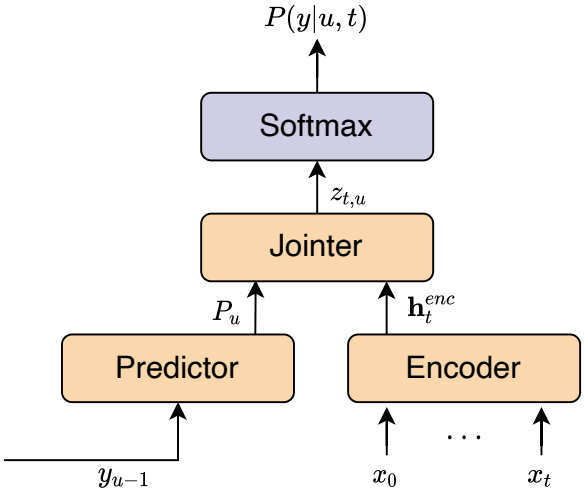


Figure 3.2: Schematic of RNN-Transducer. The figure adapted from [149].

Rao et al. [157] exploit some strategies for RNN-T architecture training and find that CTC-based encoder pre-training and language model-based predictor pre-training are significantly helpful for model performance. Besides, building on sub-word units can obtain further improvement than the model building on graphemes. Zhang et al. [216] present a Transformer-based [184] RNN-T model, in which the Transformer modules are used to learn representations from speech signals and text sequences. In addition, different window sizes on the right context are explored to enable the model to work in streaming scenarios with low latency and reliable accuracy. Inspired by Jasper and QuartzNet, Han et al. [68] propose ContextNet by introducing a fully convolutional encoder into RNN-T architecture. Since CNN has weaker receptivity on long context than LSTM and Transformer, the squeeze-and-excitation (SE) layer is integrated into ContextNet to enhance long-distance dependence. Since CNN is good at capturing local information and Transformer performs well on global context representation, Gulati et al. [64] propose Conformer which combines CNN with self-attention to concurrently learn local and global features. Experimental results show significant improvement on the LibriSpeech test and test-other set.

### 3.2.3 Attention-based Encoder-Decoder Models

Another branch of end-to-end system is the attention-based encoder-decoder architecture. As shown in Figure 3.3, the encoder directly maps input acoustic features into a latent space vector $\mathbf{h}_t^{enc}$. Then an attention mechanism iteratively weighs the contribution for the current decoder output with the inputs from each encoder

time step $\mathbf{h}_t^{enc}$ and attention output from last step $\mathbf{h}_{u-1}^{att}$. The decoder generates one target vector $P(y_u|y_0, y_{u-1}, x)$ per step by conditioning on the weighted content vector $\mathbf{h}_u^{dec}$ produced by the attention mechanism and the output $y_{u-1}$ from the last decoder step in an auto-regressive fashion. Different from CTC or RNN-T architectures, the attention mechanism dynamically bridges and aligns the encoder and decoder vectors and decides how many encoder time steps map to the decoder step to solve the variable sequences mapping issue.

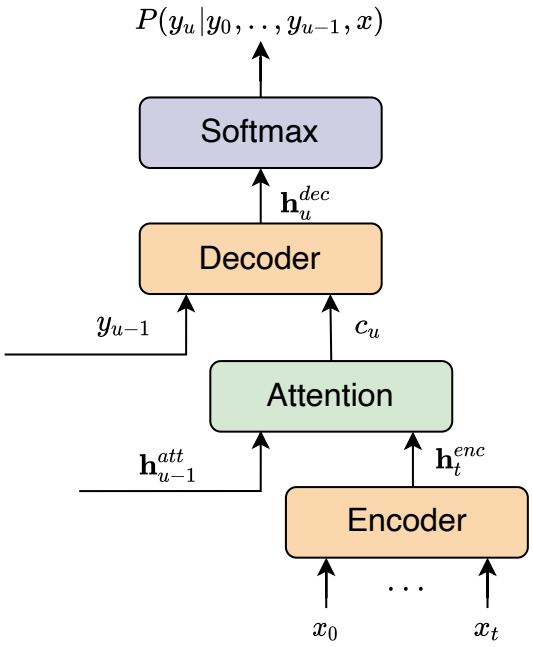


Figure 3.3: Schematic of attention-based ASR architectures. The figure adapted from [149].

Inspired by the attention success in machine translation [13] and handwriting synthesis [60], Chorowski et al. [34] transfer the attention-based recurrent networks to speech recognition and achieve competitive results on TIMIT benchmark with conventional methods. Adapted location-awareness attention mechanism taking the previous position information into account is used to alleviate the model weakness on long utterances. Based on previous attention-based ASR systems that model on phonemes, Chan et al. [25] propose LAS (listen, attend and spell) to directly map acoustic inputs to transcription in a complete end-to-end way. The LAS system comprises two components: a listener which is a pyramidal RNN used to transform speech signals into latent speech representations; a speller which converts the latent space vectors into characters/words. The authors conduct different

experiments to investigate the efficacy of each component. The results show that the pyramidal RNN structure is critical for model convergence and good performance, and the attention mechanism prevents model overfitting on the training set. In the meanwhile, Bahdanauy et al. [14] show that attention-based models can implicitly learn better context information/LM than CTC and conventional models. Experiments on large vocabulary continuous speech recognition (LVCSR) reveal the promising performance when no external lexicons or LM are used. To reduce attention computational complexity and learn more robust alignments, local monotonic attention [181], full-sequence attention [150], time-restricted self-attention [148], multi-channel attention [20], online attention [52], multi-stream self-attention [67] and triggered attention [133] are proposed to fit more for the left-to-right nature in speech recognition, and greatly promote the development of attention-based end-to-end methods.

The attention-based ASR models usually outperform CTC-based models, however they are relatively more data-hungry. In addition, they are hard to apply in streaming tasks since the attention mechanism requires long context information.

### 3.2.4 Hybrid CTC/Attention Architectures

To fully incorporate the merits of CTC and attention models, recently, Kim et al. [89] propose to jointly train CTC and attention-based approaches in a multi-task learning fashion by sharing one encoder. The evaluation on WSJ and CHiME-4 noisy speech shows the hybrid architecture can efficiently speed up the convergence process and learn more robust alignment between input frames and output sequences. Hori et al. [78] extend the hybrid CTC/attention method with a joint decoding algorithm by rescoring or combining the probabilities from both objective functions. Then, a monotonic chunk-wise attention [124] and transformer-based encoder [123] are utilized to enable the hybrid CTC-attention model working in online streaming tasks.

The hybrid CTC/attention architecture is becoming more and more popular. There has been a trend to unify the streaming and non-streaming architecture into one model with this architecture. In this thesis, we would explore the problem of OOV words with a hybrid CTC/attention architecture in Chapter 8.

## 3.3 Lip-to-Speech Reconstruction

In recent years, researchers have investigated a variety of approaches on speech reconstruction from silent videos. Compared with lipreading that converts visual sequences into text, speech construction can not only recover semantic information but also some prosodic information, like emotions, which can greatly benefit for the understanding of natural language. We only review the neural network methods in this chapter.

Le Cornu et al. [42] propose to use fully connected neural networks to estimate spectral envelope representations, for instance linear predictive coding (LPC) coefficients and Mel-filterbank amplitudes, from visual feature inputs, such as two-dimensional discrete cosine transform, followed by a STRAIGHT vocoder [208, 85], which is used to synthesize time-domain speech signals from the estimated representations. Follow-up work [105] predicts speech-related codebook entries with a classification framework to get further improvement on speech intelligibility.

Instead of using handcrafted visual features, Ephrat et al. [51] utilize CNNs to automatically learn optimal features from raw pixels and show promising results on out-of-vocabulary experiments. Subsequently, improved results are reported by Ephrat et al. [49] via combining ResNet backbone and a post-processing network on a large-scale vocabulary dataset, TCD-TIMIT [71]. Akbari et al. [4] treat the intermediate bottleneck features learned by a speech auto-encoder as training targets by conditioning on lip reading network outputs.

Kumar et al. [102] validate the effectiveness of using multiple views of faces on both speaker-dependent and -independent speech reconstruction. Vougioukas et al. [188] utilize generative adversarial networks (GAN) to directly predict raw waveforms from visual inputs in an end-to-end fashion without generating an intermediate representation of audio.

Afterwards, Prajwal et al. [151] improve the model performance with 3D CNN and skip connections. Recently, Michelsanti et al. [128] have presented a multi-task architecture to learn spectral envelope, aperiodic parameters and fundamental frequency separately, which are then fed into a vocoder for waveform synthesis. They integrate a connectionist temporal classification (CTC) [61] loss to jointly perform lip reading, which is capable of further enhancing and constraining the video encoder. However, most existing work only focuses on a speaker-dependent setting and small vocabulary or artificial grammar datasets. In Chapter 5, we evaluate our method not only on speaker-dependent experiments but also pay attention to speaker-independent and large-scale vocabulary setups.

## 3.4 Lip Reading

Lip reading, also known as visual speech recognition, is the task to predict text transcriptions from silent videos, such as mouth or face movement sequences. Research on lip reading has a long tradition. Approaches to lip reading generally fall into two categories on feature level:

a) handcrafted visual feature extraction, such as Discrete Cosine Transform [74], Discrete Wavelet Transform [147] or Active Appearance Models [176]; b) representations learned by neural networks, which has become the dominant technique for this task, for example, using convolutional auto-encoders [140], spatio-temporal convolutional neural networks [10], long short-term memory [190], or residual networks [175].

Alternatively, methods on modeling units for lip reading can be divided into word- and character level:

a) In the case of word level units, lip reading is simplified as a classification task. Word level lip reading datasets and benchmarks are built, for instance LRW [37] for English and LRW-1000 [210] for Chinese. Stafylakis et al. [175] adopt spatiotemporal convolutional networks and 2D ResNet as front end to extract visual features and bidirectional Long Short-Term Memory networks as back end to capture temporal information, and attain significant improvement. Weng et al. [199] present two separated deep 3D CNN front ends to learn features from grayscale video and optical flow inputs, respectively. Martinez et al. [120] replace recurrent neural networks widely used in past work with Temporal Convolutional Networks to simplify the training procedure. The word level methods are usually able to achieve high accuracy, however, the models disregard the interaction or co-articulation phenomenon between phonemes or words. A predefined lexicon with closed-set vocabulary is used and words are usually treated as isolated units in speech. Thereby, long-term context information, assimilation or dissimilation effects are neglected. Moreover, it is hard to recognize out-of-vocabulary words.

b) Lip reading models with character- or phoneme level mainly use methods proposed in speech recognition. Assael et al. [10] conduct end-to-end lip reading experiments on sentence level with CTC loss. Subsequently, sequence discriminative training [180] and domain-adversarial training [191] are introduced to lip reading. Chung et al. [36] collected the dataset, 'Lip Reading Sentences' (LRS) which consists of hundreds of thousands of videos from BBC television, and significantly promote the research on sentence level lip reading. Shillingford et al. [169] verify the effectiveness of large-scale data (3,886 hours of video) for training continuous visual speech recognition. Afouras et al. [1] compare the performance of recur-

rent neural networks, fully convolutional neural networks and Transformer on lip reading character recognition.

Different from the mainstream methods which directly transform videos to text, in Chapter 4 and Chapter 5, we perform lip reading experiments in a cascaded manner, in which the silent videos are firstly mapped to audio with our LipSound2 model, then text transcriptions are predicted by fine-tuning on a pre-trained speech recognition system.

## 3.5 Self-supervised Learning

As a form of unsupervised learning, self-supervised learning leverages massive unlabelled data and aims to learn effective intermediate representations with the supervision of self-generated labels. Training unlabelled data in a supervised manner relies on the pretext tasks that determine what labels and loss functions to be used. In computer vision, the pretext tasks can be predicting angles of rotated images [55], learning the relative position of segmented regions in an image [47], placing shuffled patches back [135] or colorizing grayscale input images [217]. The video-based pretext tasks can be tracking moving objects in videos [195], validating temporal frame orders [130] and video colorization [187].

Self-supervised learning is also widely used in natural language processing. It has made substantial progress recently, where diverse pretext tasks are proposed, for instance, predicting center words using surrounding ones or vice versa [129], generating the next word by conditioning on previous words in an auto-regressive fashion [22], completing masked tokens or consecutive utterances [46], recovering the order of shuffled words [103] or the permutation of rotated sentences [109].

Inspired by the strong correlation between different modalities where, for example, the audio and visual modalities are consistent semantically or happen synchronously, more and more researchers work on multi-modal or cross-modal self-supervised learning. Multi-modal self-supervised learning aims to learn joint or shared latent spaces or representations while cross-modal self-supervised learning uses one modality for the supervision of the other. Here we only review the audio-visual modalities since this is the main focus of the thesis.

Different pretext tasks are designed according to the correspondence and synchronization of audio and visual modalities, for instance, predicting whether image and audio clips correspond to enable neural networks to classify sounds [8], learning cross-modal retrieval [39], locating the sound source in an image [9], learning representations by matching the temporal synchronization [99] or spatial alignment [131]

of audio and video clips for action recognition, combining a contrastive loss and a clustering loss to learn high level semantic representations for visual events and concepts understanding [26]. In this thesis, we focus on cross-modal self-supervised learning where the corresponding audio signals are used as the supervisors of face sequence inputs.

## 3.6 Target Speech Separation

Researchers working in this field try to inform models to only concentrate on the target output utilizing auxiliary information, such as source directions [113, 144], spatial features [29], speaker identity for multi-channel [223] and single-channel [45] setups, speaker profile for both the target and competing speakers [205].

Recently, there has been a growing interest in using multimodal audio-visual methods in target speech separation. Rather than only refining a target spectrogram and reconstructing a waveform with the phase from noisy speech, Afouras *et al.* [194] use convolutional neural networks for both magnitude and phase estimation conditioning on lip regions in the corresponding video.

Furthermore, considering the fact that the visual streams may be corrupted from realistic environments, for example, when the mouth region of the speaker is occluded by a microphone, Afouras *et al.* [3] combine lip movement and self-enrolled voice representation to improve the robustness of the proposed system and to prevent the domination of the visual modality. In a similar work, Ephrat *et al.* [50] validate the effectiveness of using the whole face embedding, instead of just the lip area [194, 3], to learn the target speaker magnitude mask based on a large-scale dataset in real-world scenarios. Different from previous works focusing on time-frequency masks, Wu et al. [204] directly estimate a raw waveform in the time domain by extending the audio-only (single-modal) TasNet [115] into the audio-visual (multi-modal) domain.

Gu *et al.* [63] explore the effectiveness of using more information, i.e. speaker spatial location, voice characteristics, and lip movements, in target speech separation. A factorized attention mechanism was introduced to dynamically weigh the three kinds of additional information at the embedding level. Different from previous audio-visual works using corresponding video streams as auxiliary information, the objective of Chapter 6 in this thesis is to investigate the benefit of the pre-enrolled face image for target speech separation.

## 3.7 Learning Associations between Faces and Voices

Inspired by the finding by neuroscientists [16, 122] and psychologists [23, 165] that there is a strong relationship between faces and voices and sometimes humans can even infer what one's voice sounds like by only seeing the face, or vice versa, researchers in computer science have conducted a large number of studies on learning face and voice association that can be mainly divided into two categories: crossmodal representation and joint/shared representation.

Work on the crossmodal representation has led to the possibility of generating one modality from another, e.g. reconstructing human faces by only conditioning on speech signals. Oh *et al.* [136] design neural networks to directly map speech spectrogram to face embeddings which were pre-trained for face recognition, then decoded the predicted face representation to canonical face images with a separate reconstruction model. Wen *et al.* [198] utilize generative adversarial networks (GAN) to generate human faces from the output of a pre-trained voice embedding network. Instead of using a pre-trained network, Choi *et al.* [33] build speech and face encoders on a speech-to-face identity matching task, and train the encoders and a conditional generative adversarial network end to end to conduct face generation.

Researchers working on joint representation learning attempt to find a joint or sharing face-voice embedding space for tasks of crossmodal biometric retrieval or matching, e.g. searching a corresponding face image via a given speaker voice. Nagrani *et al.* [134] adopt a self-supervision training strategy to learn joint face and voice embeddings from videos without requiring any labelled data. Kim *et al.* [87] introduce triplet loss to learn overlapping information between faces and voices by using VGG16 [172] and SoundNet [12] for visual and auditory modality respectively. Wen *et al.* [197] propose DIMNet to leverage identity-sensitive factors, such as nationality and gender, as supervision signals to learn a shared representation for different modalities. Based on the strong association between faces and voices, in Chapter 6, we propose to utilize face embedding to guide models in tracking desirable auditory output.

## 3.8 Phonetic Knowledge Integration in Speech Recognition

We have reviewed some work on the ASR architecture and the combination of additional visual inputs. In this section, we review approaches on the integration of domain knowledge into ASR systems. There are lots of approaches focusing on incorporating domain knowledge to improve ASR performance, such as in feature engineering: Mel-frequency cepstral coefficients [43] and vocal tract length normalization [108], and in algorithm optimization: sequence discriminative training [185]. Here, we only consider studies that involve linguistic and phonetic knowledge.

Lee et al. [107] proposed automatic speech attribute transcription (ASAT) which is a new detection-based speech recognition paradigm. Compared to conventional ASR top-down paradigms, ASAT is bottom-up and coincident with the mechanism of humans perceiving and producing speech. To further improve phonological feature detection accuracy, Yu et al. [212] replaced multi-layer perceptrons by DNNs when building attribute detectors. Based on the high attribute detection precision, excellent phoneme estimate accuracy was obtained on the WSJ0 benchmark. Siniscalchi et al. [173] integrated acoustic-phonetic information into lattice rescoring. Inspired by shared phonetic knowledge among different languages, Siniscalchi et al. [174] designed a universal set of phones and used the set to improve the performance of cross-language phone recognition. Pitch accent was proposed by Ananthakrishnan et al. [7] to rescore the N-best results outputted from a standard ASR system. At present, the works integrating knowledge into ASR are mostly based on HMM hybrid architectures. Our approaches mainly focus on combining domain knowledge with neural end-to-end ASR systems in Chapter 7.

## 3.9 The Recognition of OOV Words in End-to-End ASR Models

Since the OOV problem has been explored over the last decades, plenty of approaches are proposed for conventional GMM-HMM models [167, 127] and hybrid DNN-HMM models [96, 80]. In this section, we only review methods towards end-to-end ASR architectures which have been the most promising trends in speech recognition in recent years.

Aleksic et al. [5] extend class-based LM [21, 121] by creating user-dependent small LM for contact name recognition on the application of voice command, which

is compiled dynamically based on the contact names on user's devices. Moreover, contacts insertion reward is proposed to avoid excessive bias and to balance the information between user-dependent and user-independent cases. Hori et al. [79] combine word level with character level language modeling in end-to-end architectures. With word level LM, the model can achieve better performance by learning stronger and longer context information, while character level LM is used to overcome the OOV issue that the word level LM suffers from. A similar idea is investigated by Li et al.[111], on acoustic-to-word mapping employing character level modeling units to tackle OOV issues. Williams et al. [201] leverage contextual information, for instance, user's locations, user's favorite songs, and calendar events, to partially rescore the output likelihood from sequence-to-sequence models during beam search instead of bringing an additional LM in. Since previous work does not consider errors generated by speech recognition systems when combining with external LM, Guo et al. [65] incorporate a spelling correction model into the speech recognizer training, that directly maps speech recognizer outputs to ground truth texts. The experimental results suggest that the proposed spelling correction model outperforms n-best LM rescoring and TTS data fine-tuning.

To enable on-device end-to-end speech recognition models to individually recognize new named entities, for example, the contact names on mobile phones, Sim et al. [171] compare LM biasing and acoustic model fine-tuning methods. Besides, several techniques, such as layer freezing, early stopping, and elastic weight consolidation, are investigated to suppress model overfitting during fine-tuning. Instead of using a word level LM in ASR in which a pre-defined lexicon is required, Likhomanenko et al. [112] attempt to decode acoustic models with a character level LM which is not constrained by lexicons. The lexicon-free decoder achieves better results on OOV experiments since the character level LM is naturally able to handle unseen words. To further improve the model performance on proper nouns, Zhao et al. [219] optimize the shallow-fusion method [84] (integrate an external LM into the inference of a sequence-to-sequence model) by building LM on subword level instead of at word level. Additionally, early contextual finite state transducer (FST) is proposed to avoid the proper noun candidates being pruned during the Viterbi beam search. Moreover, a common set of prefixes is utilized to avoid the contextual biasing always being active and prevent models from degrading on cases not containing OOV.

Different from most of the previous work focusing on LM post-processing which requires candidate units existing in n-best lists or decoding lattices, in Chapter 8, we tackle the root of the OOV problem and eliminate the bias in acoustic modeling

to recognize OOV words acoustically.

## 3.10 Data Augmentation with Synthetic Audio for ASR

Proper speech data augmentation can not only boost model performance but also significantly improve system robustness and generalization [97]. There are many strategies used in ASR training, for example, noise addition, pitch shifting, speed perturbation, back-translation [72] and room impulse response injection with real or simulated data[98]. More recently, a simple yet effective approach, SpecAugment [141], is proposed and achieves state-of-the-art results on LibriSpeech benchmark corpus. The basic idea for SpecAugment is randomly masking or cropping a fixed area on spectrograms in the time or frequency domain, which is able to effectively prevent model overfitting, especially for noisy conditions. Another popular method is mixing synthetic audio with real data by leveraging the advanced TTS models, like Tacotron2 [168], DeepVoice3 [146] and FlowTron [183].

Rossenbach et al. [161] compare commonly-used data augmentation strategies with the TTS audio using the texts in training set on attention-based models. The results reveal the effectiveness of TTS data in ASR system training. Laptev et al. [104] investigate the effect of augmenting data with TTS audio for low-resource speech recognition. The results outperform other systems with the same setting and semi-supervised learning methods. Besides, authors explore the influence of audio quality with different vocoders, i.e. Griffin-Lim and LPCNet [182]. Instead of just mixing the synthetic audio data during training, Rosenberg et al. [160] exploit the impact of TTS model effectiveness and diversity on ASR results. Moreover, lexical diversity is also investigated on domain adaptation experiments.

Inspired by the benefit brought by TTS data, in this thesis, we synthesize audio with text crawled from the Internet containing OOV words as the training set for new vocabulary acquisition and model adaption.

## 3.11 Summary

We provide a comprehensive review in this chapter. Firstly, we present the state-of-the-art end-to-end architectures, i.e. CTC, RNN-T and attention-based encoder-decoder models, followed by the related work on lip-to-speech reconstruction, lip reading and self-supervised learning that are the main focus of Chapter 4 and

5. Then, the recent progress on target speech separation and learning associations between faces and voices are reviewed. Finally, we present some relevant techniques on domain knowledge integration, OOV word recognition and data augmentation with synthetic audio.

# Chapter 4

# LipSound: Neural Mel-spectrogram Reconstruction for Lip Reading

## 4.1 Introduction

Recently, automatic speech recognition (ASR) has accomplished great progress, and advanced models are proposed achieving impressive performance on a variety of benchmarks [6, 34, 14, 24], reaching human parity on some tasks [206]. However, in realistic environments, the performance of ASR systems suffers from significant degradation because of environmental noise or ambient reverberation [88, 15, 70, 200].

Inspired by human bimodal perception [18] in which both visual and auditory information are used to improve the comprehension of speech, a lot of effort has been spent on lip reading to predict text transcriptions directly from visual cues and improve the robustness of ASR [145, 175, 1, 10, 36, 207]. The visual signal is invariant to acoustic noise and complementary to auditory representation [118, 170], and the visual contribution becomes more important as the acoustic speech-to-noise ratio is decreased [177].

Approaches to lip reading generally fall into two categories: (a) handcrafted visual feature extraction, in which many methods have been proposed based on visual signal processing algorithms. For instance, Discrete Cosine Transform [74], Discrete Wavelet Transform [147], Active Appearance Models [176]; (b) automatic feature extraction using neural networks. This has become the dominant technique in this task, for example, using convolutional auto-encoder [140], spatio-temporal

convolutional neural networks [10], long short-term memory [190], and residual networks [175].

Although advanced feature engineering and powerful deep neural network architectures have been proposed, lip reading still cannot be competitive with speech recognition from audio. It is mainly because the visual modality carries less relevant information for recognition than audio. Furthermore, some phonemes are visually identical but different and discriminative in audio. For example, in English the minimal pairs /b/ and /p/, where /b/ is a voiced sound and /p/ is an unvoiced sound. They are different in audio-based speech representation, while the two phonemes are modeled as the same unit in traditional lip reading systems, since /b/ and /p/ are produced with the same visually apparent lip and tongue movement.

Inspired by the success of the Tacotron2 [168] which only requires text as input to predict the Mel-spectrogram for speech synthesis, in this chapter, we propose to map the image sequences of the mouth region directly to Mel-spectrogram to reconstruct the relevant acoustic information.

Our proposed model architecture consists of two main components, (i) a recurrent encoder and decoder with an attention mechanism front-end that generates Mel-scale spectrograms from image sequences of video. Unlike end-to-end lip reading models, this component can be trained with large amounts of non-annotated video data. (ii) a lip reading back-end that maps the generated Mel-spectrogram to text directly. We conduct the evaluation of the overall model on the lip reading benchmark GRID dataset.

During training, instead of consuming the predicted output from previous time steps, we use teacher-forcing training strategy [202] to utilize the ground truth speech spectrogram as input, which is different from the traditional lip reading architectures that map the sequence of images directly to text transcriptions. Besides, the temporal dependencies between consecutive acoustic frames enable the front-end model not only to reconstruct the segmental features, for example formants, but also the supra-segmental information, for instance speaking styles. The reconstructed speech-relevant information significantly improves the performance of the lip-reading back-end. The model described in this Chapter has been published in Qu et al [154].

## 4.2 Dataset and Pre-processing

GRID [41] is a current benchmark and biggest open source lip reading dataset, which consists of in total 34000 videos from 34 speakers (16 female and 18 male) on sentence level. The sentences have a fixed 6-word structure and are generated by a restricted grammar: $command^{(4)}$ + $color^{(4)}$ + $preposition^{(4)}$ + $letter^{(25)}$ + $digit^{(10)}$ + $adverb^{(4)}$, where the superscript means the number of candidate words, for example "Set red by Z two now". The category details are listed in Table 4.1.

Table 4.1: GRID dataset word categories

| **Categories** | **Candidate Words** |
|---|---|
| Command | bin, lay, place, set |
| Color | blue, green, red, white |
| Preposition | at, by, in, with |
| Letter | A, . . . , Z (W excluded) |
| Digit | zero, . . . , nine |
| Adverb | again, now, please, soon |

GRID consists of 34 speakers, but there are only 33 speakers' videos available with a total of 32669 sentences. We randomly select 255 sentences from each speaker used for evaluation. The remaining sentences are divided into training and development sets with a ratio of 9:1. Both training and test sets contain samples of all speakers, leading to speaker-dependent results.

All videos are 3 seconds long and with a frame rate of 25 fps (total 75 frames). All frames are converted from original RGB color to gray images. To collect training data for the Mel-spectrogram generator, we use FFmpeg to extract audio and downsample to 16kHz. We use Dlib's Python bindings [90] to detect 68 facial landmarks which are used to crop an area of 100x50 pixels around the mouth from each frame. Then we regularize the pixel value to [-1, 1] and normalize the mean value and standard deviation of all images to 0 and 1 respectively. The input shape of the Mel-spectrogram generator is 75x512, where 75 (3x25) is the number of frames extracted from videos and 512 is the dimension of each mouth area compressed by PCA.

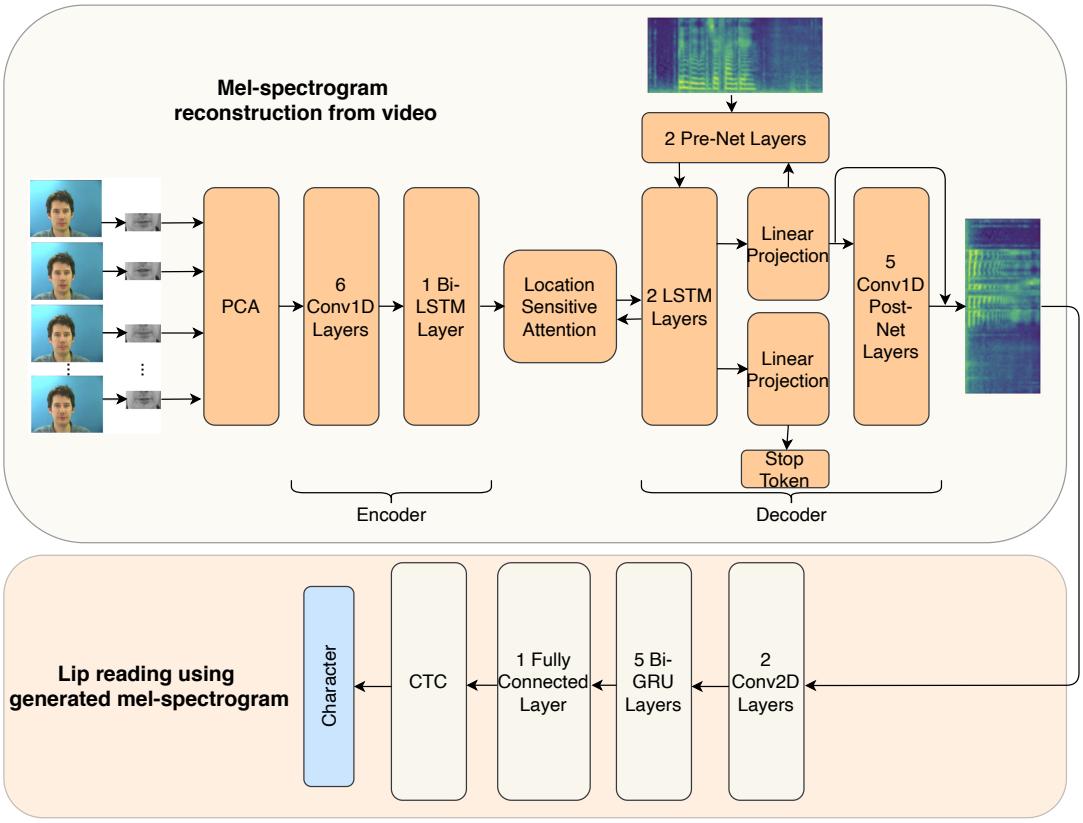


Figure 4.1: LipSound model architecture. The front-end (top part) is used for Mel-spectrogram reconstruction and the back-end (bottom) is used for character recognition. Together, they perform lip reading.

## 4.3 Model Architecture

### 4.3.1 Front-end: Mel-spectrogram Generator

The Mel-spectrogram generator is inspired by the Tacotron2 which only uses character embedding and corresponding speech waveform as input for training, without requiring any linguistic, handcrafted features or domain expertise knowledge. When combining it with the Wavenet vocoder [178], the Tacotron2 achieves high-quality sounds which can be comparable to natural human speech. In the Mel-spectrogram generator, we replace the character sequences with video representation as the model's inputs and remove the word embedding layer since it is unnecessary in our case.

As shown in Figure 4.1 (top part), our Mel-spectrogram generator consists of an encoder, a decoder and an attention mechanism, which is similar to the Tacotron2 system. The encoder compresses the image sequences from the mouth area into

latent vectors and the attention mechanism learns to align the encoder and decoder time steps and focuses on the most relevant information for the current step. Finally, the decoder consumes the attention context vector and all the previous information to predict the Mel-spectrogram step by step.

The images of size 100x50 are compressed to 512 dimensions by PCA and then directly fed into 6 layers of a 1D convolutional neural network with 512 filters and a kernel size of 5. These CNN layers have a similar function as the N-grams used in natural language processing to capture the temporal information in multiple adjacent frames. Each CNN layer is followed by batch normalization and rectified linear unit (ReLU) activation functions. The outputs from CNN layers are consumed by one bi-directional LSTM layer.

Attention mechanisms have become standard in encoder and decoder architectures since they reduce computational complexity and let the model focus on the most relevant information. We use location-sensitive attention [30] to direct the information flow from the encoder to the decoder, which focuses both content and location information to predict the next decoding time step and yields smoother alignments.

The decoder consists of two LSTM layers, 2 fully connected layers (pre-net) and 1 linear projection layer, which evaluates an output Mel-spectrogram one frame at a time. Only during training the ground truth Mel-spectrogram extracted from the corresponding speech waveforms is fed into the pre-net layers. We use the predicted Mel-spectrogram from previous time steps when inferring. Then the output from pre-net is concatenated with attention context vectors as the inputs of the following 2 LSTM layers. The output is used to generate Mel-spectrogram by one sigmoid projection layer at this time step. In the meantime, the output from the LSTM layers is also consumed by another sigmoid linear layer to predict the stop token. This is useful for the inference phase since all sentences in the training set are zero-padded to have the same dimensionality. The stop token predicts when to terminate decoding and avoids always generating the same duration and silence padding for sentences of short duration. Finally, to further improve the Mel-spectrogram quality, the predicted Mel-spectrogram is fed into 5 convolutional layers with residual connections, named post-net. Both the Mel-spectrogram output from the linear layer and post-net are used for lip reading back-end model training.

### 4.3.2 Back-end: Lip Reading System

We use DeepSpeech 2 [6] ASR system as back-end module to transcribe spectrogram into text, as shown in Figure 4.1 (bottom part), which begins with two layers of 2D convolutions, followed by five layers of gated recurrent units (GRU) [31] and a fully connected output layer. Finally, we use the connectionist temporal classification (CTC) loss [61] to calculate the difference between the predicted transcriptions and the ground truth.

## 4.4 Experiments

In this section, we introduce two audio based speech recognition systems as gold standard and conduct experiments to reconstruct Mel-spectrogram from videos with the GRID dataset. The predicted spectrograms are evaluated on lip reading tasks.

### 4.4.1 Setups for Mel-spectrogram Generator and Lip Reading

The feature prediction experiments are conducted on a single NVIDIA 1080Ti GPU card with a fixed mini-batch size of 30. We used the Adam optimizer [92] with an initial learning rate of 0.001. We found annealing the learning rate with a value of 1.1 after every 50000 iterations can achieve the best results.

The input features for our lip reading systems are Mel-spectrogram. The back-end neural networks are trained with the CTC loss function, using the stochastic gradient descent optimization strategy along with a mini-batch of 30 utterances per batch. We use 40 epochs and pick the model that performs best on the development set to quantify on the test set. Learning rates are chosen from [1e-4, 6e-4] and a learning rate annealing algorithm is used after each epoch. Batch normalization is used to optimize models and accelerate training on hidden layers. All architectures described in this chapter do not use any language models.

### 4.4.2 Audio Gold Standard Models for Lip Reading

We use word error rate (WER) and character error rate (CER) as the evaluation metric. We establish two strong audio gold standards in this work. Audio gold standard 1 is trained from scratch using only the Mel-spectrogram features extracted directly from the original training set. Audio gold standard 1 achieves

better performance, with 0.8% CER and 2.1% WER, than all lip reading systems [10, 36, 207] that only use the visual modality as input. This result also verifies that the speech modality contains more useful information for recognition than the visual modality.

To further improve the performance, we use the 960 hours LibriSpeech [138] training set to pretrain the audio gold standard model. LibriSpeech is a large open source speech corpus and a widely-used speech recognition benchmark. The pretrained acoustic model achieves 11.43% WER on the LibriSpeech clean evaluation set after 13 epochs. After fine-tuning 18 epochs on the pretrained LibriSpeech acoustic model, the audio gold standard 2 gets significant improvement with 0.2% CER and 0.6% WER on the GRID test set.

### 4.4.3 Evaluation Metric

We use the word error rate (WER) to evaluate model performance. WER quantifies how many elementary operations are required to transform the generated output sequence of the network into the correct target sequence. It is calculated as follows:

$$WER = \frac{S + D + I}{N} \tag{4.1}$$

where $S$ is the number of substitutions, $D$ is the number of deletions and $I$ is the number of insertions. $N$ is the total number of words in the reference.

## 4.5 Results and Discussion

### 4.5.1 Results of Mel-spectrogram Reconstruction

As shown in Figure 4.2, we visualize the original Mel-spectrogram (top row) extracted from the corresponding audio as references. We also show the Mel-spectrogram samples (bottom row) generated from our feature prediction front-end.

Figure 4.2 (a) shows two correctly generated samples. As shown in the figure, the Mel-spectrogram predicts highly similar details to the original one, with similar starting and ending time, low frequency and formants. We tend to attribute this performance to the good alignment learned by the attention mechanism between the encoder and decoder time steps. But the high frequencies seem fuzzy and are not as clear as in the original Mel-spectrogram.

We transform the generated Mel-spectrogram back to the waveform using the Griffin Lim algorithm [62] with 50 iterations. The Griffin Lim algorithm is widely

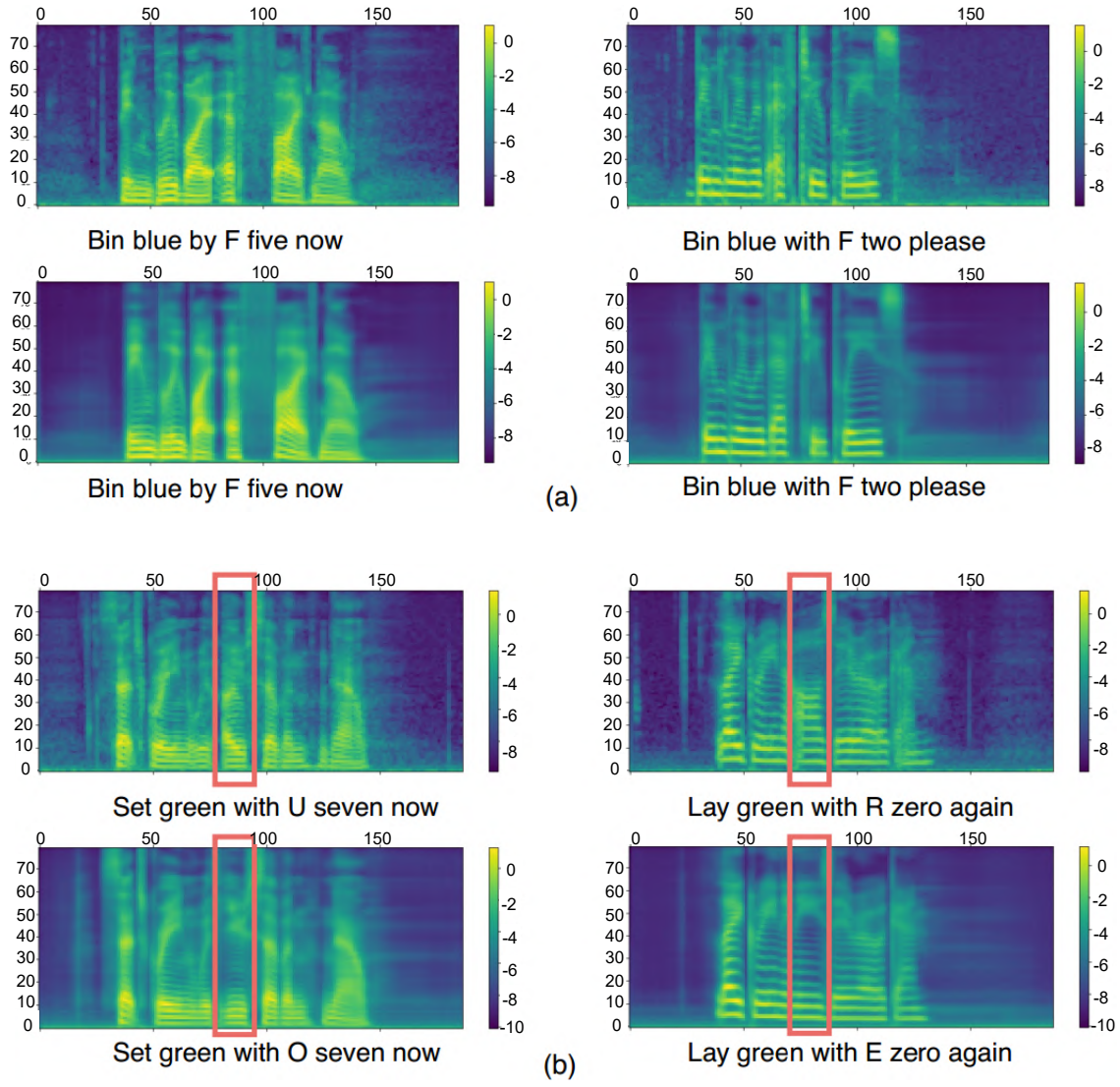


Figure 4.2: The comparison of real (top row) and generated (bottom row) Mel-spectrogram. (a) correct generation. (b) generated Mel-spectrogram with word substitution (as marked with red rectangles).

used to restore phase information for waveform reconstruction. We can verify that the Mel-spectrogram generator can learn different speakers' voices. But by directly hearing the generated sounds, we find that some of the isolated letters have been substituted by another letter, as shown in Figure 4.2 (b). For example, the letter 'O' in the sentence 'Set green with O seven now' is replaced by 'U' and the letter 'R' in 'Lay green with R zero again' is replaced by 'E'. However, the other words are inferred correctly. Unlike the words with multiple letters, the isolated letters are independent of context. For the same left and right context, for example

'with+*+seven' where * means A-Z (excluding W), there are 25 possibilities. It is difficult for the decoder to infer a correct letter using the same context information.

Besides, after checking the original sounds, we found that speakers tend to have a short pause before producing isolated letters. The short silence affects attention alignment radically since silence parts do not contain any semantic information or any content dependency. Figure 4.3 shows the influence of silence on alignment. The beginning and end of the sentence is silence, causing a divergent alignment instead of an intensive yellow line. More audio samples are available on `https://soundcloud.com/user-612210805/sets/video-to-mel`.

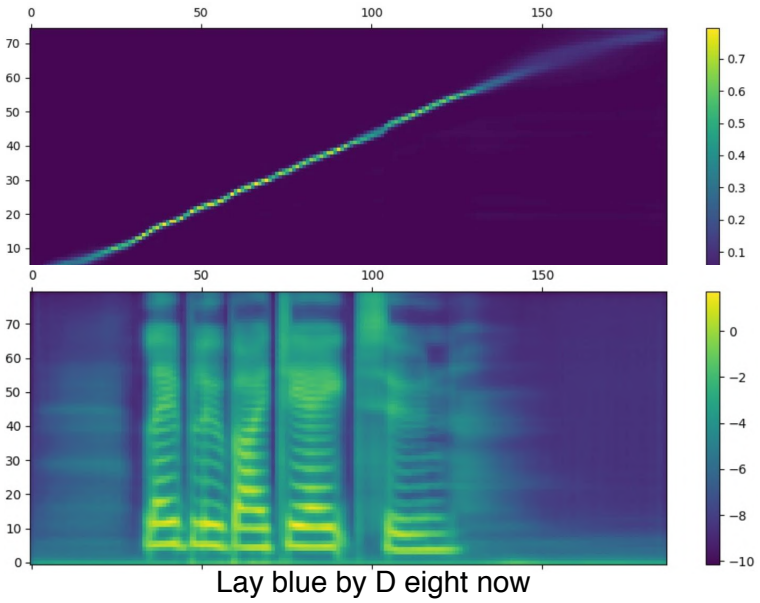


Figure 4.3: Alignment between encoder and decoder time steps. Top: the attention mechanism alignment curve (yellow diagonal line). Bottom: Mel-spectrogram generated from post-net.

### 4.5.2 Results of Lip Reading

As expected, the gold standard models trained on audio outperform the models [10, 36, 207] trained only on visual information. This shows again that speech carries more useful information for speech recognition than the visual modality.

Table 4.2 shows the comparison between our proposed LipSound and previous works. All cited works [10, 36, 207] predicted text transcriptions from videos directly. LipNet [10] is trained in an end-to-end fashion on a sentence level which makes use of spatio-temporal convolutions and CTC and achieves 1.9% CER and

Table 4.2: CER and WER comparison on the GRID lip reading dataset. All cited works use visual information as model inputs. Audio gold standard 1 is trained on the GRID audio dataset. Audio gold standard 2 is pre-trained on the LibriSpeech acoustic model. NoLM: no language models are used.

| **Model** | **CER (%)** | **WER (%)** |
|---|---|---|
| ***Audio:*** | | |
| Gold standard 1-NoLM | 0.811 | 2.053 |
| Gold standard 2-NoLM | 0.180 | 0.564 |
| ***Visual:*** | | |
| LipNet-NoLM [10] | 2.0 | 5.6 |
| LipNet [10] | 1.9 | 4.8 |
| WAS [36] | - | 3.3 |
| LCANet [207] | 1.3 | 2.9 |
| LipSound-NoLM | 1.532 | 4.215 |
| LipSound with pretrain-NoLM | **0.843** | **2.525** |

4.8 WER%. The WAS [36] network utilizes an encoder-decoder with an attention architecture and pretrains on the Lip Reading Sentences dataset which is a large-scale dataset for audio-visual speech recognition. The WAS model yields 3.3% WER on the GRID evaluation set. The LCANet networks introduced a cascaded attention-CTC decoder to further improve the performance and achieved 1.3% CER and 2.9% WER.

Our model trained on the visually generated Mel-spectrogram achieves 1.532% CER and 4.215% WER. To further improve the accuracy, we fine-tune the lip reading model on the pretrained LibriSpeech model with updating all parameters. After 15 epochs, we get better than state-of-the-art performance with 0.843% CER and 2.525% WER.

Table 4.3 lists the comparison between the ground truth and the predicted text transcriptions. As reported [10], the frequently confused phoneme pairs are (d, t) and (b, p), while in our results, the most frequent errors are letter substitutions, such as (A, H) where 'A' is substituted by 'H'. This indicates that our Mel-spectrogram front-end has reconstructed the lost information in the visual representation, while it needs to make guesses for phonemes that are easily confused visually.

Table 4.3: Comparison between ground truth and predicted sentence by our lip reading system. Mistaken words are underlined.

| **Ground truth** | **Predicted sentencs** |
|---|---|
| Lay blue in A seven please | Lay blue in H seven please |
| Place red in N zero soon | Place red in A zero soon |
| Lay red in O seven please | Lay red in I seven please |
| Bin green by L seven now | Bin green by S seven now |
| Lay green at X six soon | Lay green at X six sooen |
| Set blue in R three please | Set blue in R three pleae |
| Lay white at C seven now | Lay white it C seven now |

## 4.6 Summary

We proposed a novel architecture, LipSound, for lip reading in which an encoder-decoder architecture with attention mechanism is used to reconstruct Mel-spectrogram from the image sequences of videos directly. The encoder encodes source image sequences into a context vector, and the decoder decodes the context vector to predict a target Mel-spectrogram. The attention mechanism learns to align the encoder and decoder time steps and to concentrate on the most relevant information. The lip reading back-end consumes the generated Mel-spectrogram representation to predict text transcriptions. The speaker-dependent evaluation results on the GRID benchmark dataset demonstrate that our system outperforms state-of-the-art performance of existing models.

# Chapter 5

# LipSound2: Self-supervised Pre-training for Lip-to-Speech Reconstruction and Lip Reading

## 5.1 Introduction

In Chapter 4, we proposed LipSound [154] to directly map visual sequences to low level speech representation, i.e. Mel-spectrogram, which is inspired by audio-visual self-supervised representation learning. By leveraging the natural co-occurrence of audio and visual streams in videos without requiring any human annotations, or treating one modality as the supervision of the other, self-supervised representation learning has received substantial interest, for example, learning representations by matching the temporal synchronization [99] or spatial alignment [131] of audio and video clips for action recognition. In this chapter, we further explore to what extent the large-scale crossmodal self-supervised pre-training can benefit speech reconstruction in generalizability (speaker-independent) and transferability (Non-Chinese to Chinese).

Inspired by human bimodal perception [18] in which both sight and sound are used to improve the comprehension of speech, a lot of effort has been spent on speech processing tasks by leveraging visual information, for example, integrating simultaneous lip movement sequences into speech recognition [36, 1], guiding neural networks in isolating target speech signals with a static face image for speech separation [155, 38] and grounding speech recognition with visual objects and scene information [125, 66]. Multi-modal audio-visual methods achieve significant improvement over single modality models, since the visual signals are invariant to

acoustic noise and complementary to audio representations [118]. Moreover, the visual contribution becomes more important as the acoustic signal-to-noise ratio is decreased [170].

In most approaches, the visual information is mainly used as auxiliary input to complement audio signals. However, in some circumstances, the auditory information may be absent or extremely noisy, which motivates the use of speech reconstruction. Speech reconstruction aims to generate both intelligible and qualified speech by only conditioning on image sequences of talking mouths or faces. Generating intelligible speech from silent videos enables many applications, e.g. a silent visual input method on mobile phones for privacy protection in public areas; communication assistance for patients suffering laryngeal disorders, like laryngectomy; surveillance video understanding when only visual signals are available; enhancement of video conferences or far-field human-robot interaction scenarios in a noisy environment; non-disruptive user intervention for autonomous vehicles.

It is challenging to reconstruct qualified and intelligible speech from only mouth or face movements, since human speech is produced by not only externally observable organs, like lips and tongue, but also internally invisible ones which are difficult to capture in most cases [54], for instance, vocal cords and pharynx. Consequently, it is hard to infer fundamental frequency or voicing information controlled by these organs. Moreover, some phonemes are acoustically discriminative but not easy to distinguish visually since the phonemes share the same places of articulation but with different manners of articulation [119], for example, /v/ and /f/ in English are both fricatives and look the same on lip and teeth movements but are different on the vibration of vocal cords (voiced vs unvoiced) and the attribute of aspirate (unaspirated vs aspirated) which are not visible in most video recordings. Hence, predicting human voices from appearance is still a challenging task [57].

In recent years, there has been a growing interest on speech reconstruction and variant methods have been proposed. A possible technique is to run lip reading (video-to-text) and text-to-speech (TTS) systems in cascade but the lip reading performance is still unsatisfactory and the error is being propagated to TTS. Alternatively, other researchers directly estimate speech representations, for example, linear predictive coding [51], bottleneck features [4], and Mel-scale spectrograms [154], from videos, followed by a vocoder used to transform intermediate representations to audio, for instance, STRAIGHT [85] and WORLD vocoder [132]. In contrast, the information of speaker identity and speaking styles can be relatively preserved. However, most existing work only focuses on speaker-dependent settings with a small vocabulary or artificial grammar dataset, or even builds one

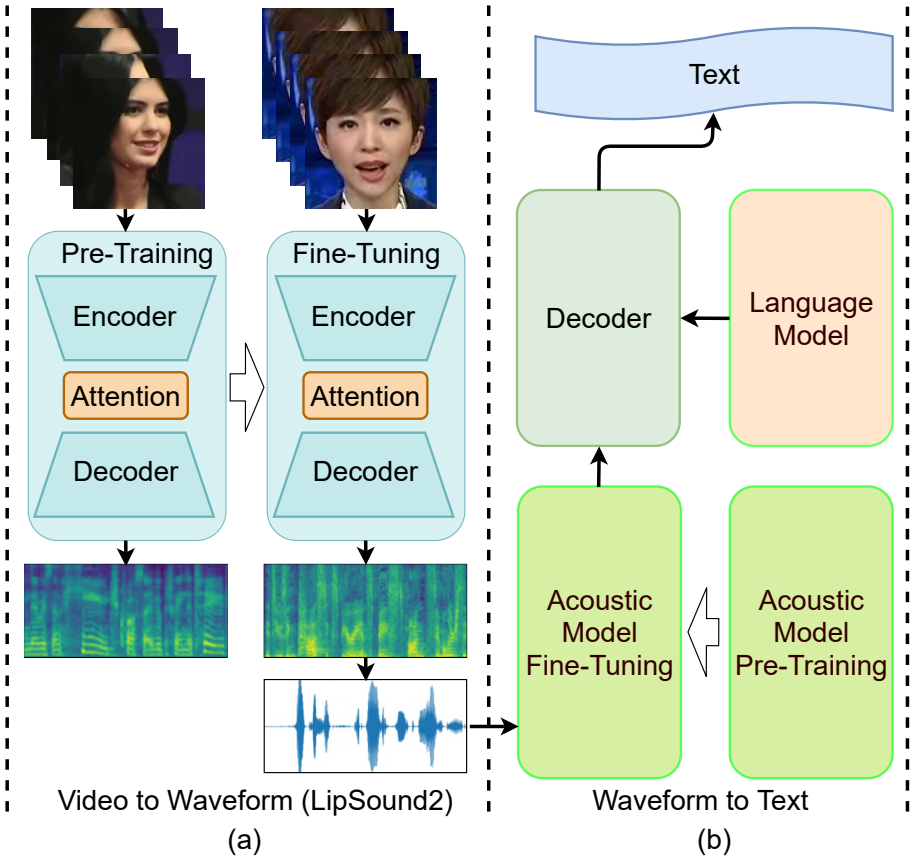


Figure 5.1: Pipeline of video to waveform generation and waveform to text transformation.

model for each individual speaker, which does not meet the requirements in realistic scenarios.

In this chapter, we upgrade our LipSound architecture by replacing 1DCNN with 3DCNN blocks (Conv 3D + Batch Norm + ReLU + Max Pooling + Dropout) to enable the model to directly learn stable representations from raw pixels and using location-aware attention mechanism to make the alignments between encoder and decoder more robust to non-verbal areas. Moreover, we replace the Griffin-Lim algorithm [62] with a neural vocoder to smoothly generate waveforms and voices. As shown in Fig. 5.1 (a), our approach is firstly pre-training the Lipsound2 model on a large-scale multi-lingual audio-visual corpus (VoxCeleb2) to map silent videos to Mel-spectrogram, then fine-tuning the pre-trained model on specific domain datasets (GRID, TCD-TIMIT and CMLR), followed by a neural vocoder (WaveGlow [152]) to reconstruct estimated Mel-spectrogram to waveforms. Lip reading (video-to-text) experiments are performed by fine-tuning the generated audio on a pre-trained acoustic model (Jasper [110]) in Fig. 5.1 (b).

The main contributions of this chapter are:

1. We propose an auto-regressive encoder-decoder with attention architecture, LipSound2, to directly map silent facial movement sequences to Mel-scale spectrograms for speech reconstruction, which does not require any human annotations.

2. We explore the model generalizability on speaker-independent and large-scale vocabulary datasets which few studies have focused on, and we achieve better performance on speech quality and intelligibility in the speech reconstruction task.

3. To the best of our knowledge, no previous research has investigated Chinese speech reconstruction in speaker-dependent and -independent cases.

4. By leveraging the large-scale self-supervised pre-training on LipSound2 and the advanced Jasper speech recognition model, our cascaded lip reading system outperforms existing models by a margin on both English and Chinese corpora.

This chapter is organised as follows. Section 5.2 gives the model details, followed by the description of datasets and evaluation metrics in Section 5.3. Experimental results and discussion are presented in Section 5.4. The model described in this Chapter is under review of the Journal of IEEE Transaction on Neural Networks and Learning Systems (Qu et. al. [156]).

## 5.2 Model Architecture

Fig. 5.2 shows the LipSound2 model architecture. We split the video clips into an audio stream used as training target and a visual stream used as model input. The system consumes the visual part to predict the audio counterpart in a self-supervised fashion. The proposed architecture is composed of an encoder-decoder and an attention model to map the soundless visual sequences to the low level acoustic representation, Mel-scale spectrograms. Advantages are that, in contrast to directly predicting raw waveform, working with Mel-spectrogram not only reduces computational complexity but also easily learns long-distance dependence. Model details are listed in Table 5.1. Then a pre-trained neural vocoder, WaveGlow, follows to reconstruct raw waveform from the generated Mel-spectrogram.

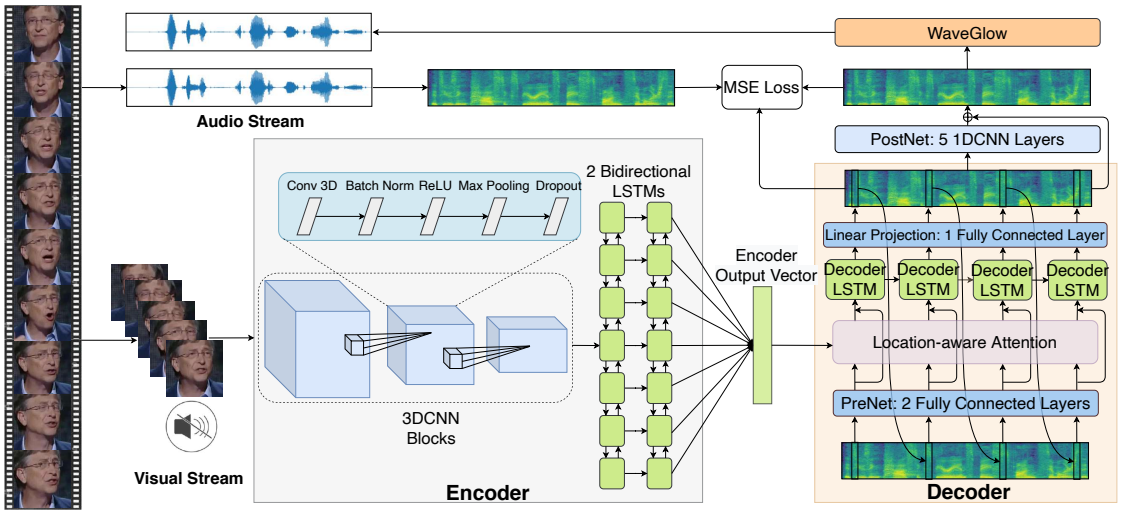


Figure 5.2: The architecture of LipSound2. The video is split into visual and acoustic streams. The face region which is cropped from the silent visual stream is used as the model input. The acoustic spectrogram features extracted from the counterpart audio stream are used as the training target. The weighted attention content vector which is generated by multiplying the encoder output and location attention weights is fed into the decoder to produce the target spectrogram frame by frame. During training, the ground truth spectrogram frames are utilized to accelerate convergence, while, during inference, the outputs from previous steps are used.

### 5.2.1 Encoder

The multi-Task CNN (MTCNN) [215] is used to detect face landmarks from raw videos. We crop only the face region ($112 \times 112$ pixels) and smooth all frame landmarks, since low-resolution videos or profile faces lead to detection failures sometimes and landmark smoothing can eliminate frame skip in adjacent images. The cropped face sequences are then fed into 3D CNN blocks and each block is based on a 3D CNN, Batch Normalization, ReLU activation, Max Pooling and Dropout, as shown in Fig. 5.2. Then two bidirectional LSTM layers follow which capture the long-distance dependence from the left and right context.

### 5.2.2 Location-sensitive Attention

We use location-aware attention [34] to form a bridge between the encoder and the decoder. As shown in Fig. 2.7, the outputs from the encoder and attention LSTM cell are mapped to a fixed dimension in hidden space via memory FC and

query FC respectively. The location convolutional layer takes the concatenation of last step attention weights and the sum of previous steps as inputs to be aware of location information and predicts the current step values, followed by the location FC. The sum of outputs from memory FC, query FC and location FC is processed by Tanh activation and weight FC, followed by a softmax function used to normalize the attention values to [0, 1]. The attention content vector can be obtained by multiplying the encoder output by the normalized attention weights. The concatenation of the attention content vector and the attention LSTM cell output will be fed into the decoder LSTM.

### 5.2.3 Decoder

The decoder module consists of one unidirectional LSTM layer and one linear projection layer. The decoder LSTM consumes the attention content vector and the output from attention LSTM to generate one frame at a time. Subsequently, the linear projection layer maps the decoder LSTM outputs to the dimension of the Mel-scale filter bank. During training, we use ground truth Mel-spectrogram frames as PreNet input and during inference, the predicted frames from previous time steps are used. Since the decoder only receives past information at every time step, after decoding, five Conv1D layers (PostNet) are used to further improve the model performance by smoothing the transition of adjacent frames and using future information which is not available when decoding.

### 5.2.4 Training Objective

The loss function is the sum of two mean square errors (MSE), as shown in Eq. (5.1), i.e. the MSE between the decoder output $O_{dec}$ and the target Mel-spectrogram $M_{tar}$ and the MSE between the PostNet output $O_{post}$ and the target Mel-spectrogram.

$$Loss = MSE(O_{dec}, M_{tar}) + MSE(O_{post}, M_{tar}) \tag{5.1}$$

### 5.2.5 WaveGlow

We use WaveGlow [152] which combines the approach of the glow-based generative model [93] and the architecture insight of WaveNet [137] to transform the estimated Mel-spectrogram back to audio. WaveGlow abandons auto-regression [137] and speeds up the procedure of waveform synthesis in high quality and resolution. We

Table 5.1: Configuration of LipSound2 encoder, decoder, attention and PostNet.

| Layer | Kernel | Stride | Padding | Channels/Nodes |
|---|---|---|---|---|
| **Encoder** | | | | |
| Conv3D 1 | $5 \times 3 \times 3$ | $[1, 2, 2]$ | $[2, 0, 0]$ | 32 |
| MaxPool3D | $1 \times 2 \times 2$ | $[1, 2, 2]$ | $[0, 0, 0]$ | - |
| Conv3D 2 | $5 \times 3 \times 3$ | $[1, 2, 2]$ | $[2, 0, 0]$ | 64 |
| MaxPool3D | $1 \times 2 \times 2$ | $[1, 2, 2]$ | $[0, 0, 0]$ | - |
| Conv3D 3 | $5 \times 3 \times 3$ | $[1, 1, 1]$ | $[2, 0, 0]$ | 128 |
| MaxPool3D | $1 \times 2 \times 2$ | $[1, 2, 2]$ | $[0, 0, 0]$ | - |
| BiLSTM1 | - | - | - | 128 |
| BiLSTM2 | - | - | - | 128 |
| **Attention** | | | | |
| Attention LSTM | - | - | - | 1024 |
| Query FC | - | - | - | 128 |
| Memory FC | - | - | - | 128 |
| Location Conv1D | 31 | 1 | 15 | 32 |
| Location FC | - | - | - | 128 |
| Weight FC | - | - | - | 1 |
| **Decoder** | | | | |
| PreNet FC 1 | - | - | - | 512 |
| PreNet FC 2 | - | - | - | 256 |
| Decoder LSTM | - | - | - | 1024 |
| Linear Projection FC | - | - | - | 80 |
| **PostNet** | | | | |
| Conv1D 1 | 5 | 1 | 2 | 512 |
| Conv1D 2 | 5 | 1 | 2 | 512 |
| Conv1D 3 | 5 | 1 | 2 | 512 |
| Conv1D 4 | 5 | 1 | 2 | 512 |
| Conv1D 5 | 5 | 1 | 2 | 80 |

train WaveGlow from scratch using the same settings as original work [152] but in 16k sampling rate on the LJSpeech dataset [81] to meet the requirement of following up ASR models. To our surprise, the WaveGlow model that is trained with only one female voice can effectively generalize to any unseen voices and

stably perform waveform reconstruction.

### 5.2.6 Acoustic Model and Language Model

The Jasper [110] speech recognition system which is a fully convolutional architecture trained with skip connections and CTC loss is adopted to directly predict characters from speech signals. We pretrain the Jasper DR 10x5 model[1] on 960h LibriSpeech and 1000h AISHELL-2 corpora, which achieves 3.61% WER and 10.05% CER on the development set for English and Chinese, respectively. Beam search is utilized to decode the output character possibilities from Jasper and a 6-gram KenLM [73] language model[2] into grammatically and semantically correct words on sentence level.

## 5.3 Experiments

### 5.3.1 Dataset

All datasets used in this chapter are summarized in Table 5.2 and random frames from audio-visual ones are presented in Fig. 5.3. VoxCeleb2 is a large-scale audio-visual corpus, extracted from YouTube videos, containing over one million utterances and more than 6k different speakers from around 145 nationalities and languages. It includes noisy and unconstrained conditions, specifically, the audio stream may be recorded with background noise, such as laughter and room reverberation, and the vision part may contain variable head poses (e.g. frontal faces and profile), variable lighting conditions and low image quality, while the GRID and TCD-TIMIT datasets are in controlled experimental environments with fixed frontal face angle and clean background in audio and vision. It is worth to mention that the GRID dataset is designed to contain only a fixed 6-word structure and all sentences are generated by a restricted artificial grammar: *command + color + preposition + letter + digit + adverb*, for example, set blue in Z three now. CMLR (Chinese Mandarin Lip Reading) is collected from videos by 11 hosts of the Chinese national news program *News Broadcast*, which contains frontal faces and covers a large amount of Chinese vocabulary. We firstly pretrain LipSound2 on VoxCeleb2, then fine-tune the model on GRID, TCD-TIMIT and CMLR respectively for video to Mel-spectrogram reconstruction.

[1] https://nvidia.github.io/OpenSeq2Seq/html/speech-recognition.html
[2] https://github.com/PaddlePaddle/DeepSpeech

LibriSpeech and AISHELL-2 are the current largest open-source speech corpora and widely-used speech recognition benchmarks for English and Chinese, respectively. LibriSpeech is derived from audiobooks, containing 460h of clean speech and 500h of noisy speech. AISHELL-2 consists of 1000h different domain speech, for instance, voice command and smart home scenario, and includes various accents from different areas of China. We use LibriSpeech and AISHELL-2 to pretrain the Jasper acoustic model to boost the performance of waveform-to-text transformation. The generated speech on GRID, TCD-TIMIT and CMLR is used for further fine-tuning to perform lip reading (video-to-text) experiments.

The LJ Speech dataset with only one female voice is especially designed for speech synthesis tasks, which is used for WaveGlow training, in this chapter, to transform Mel-spectrogram back to waveforms.

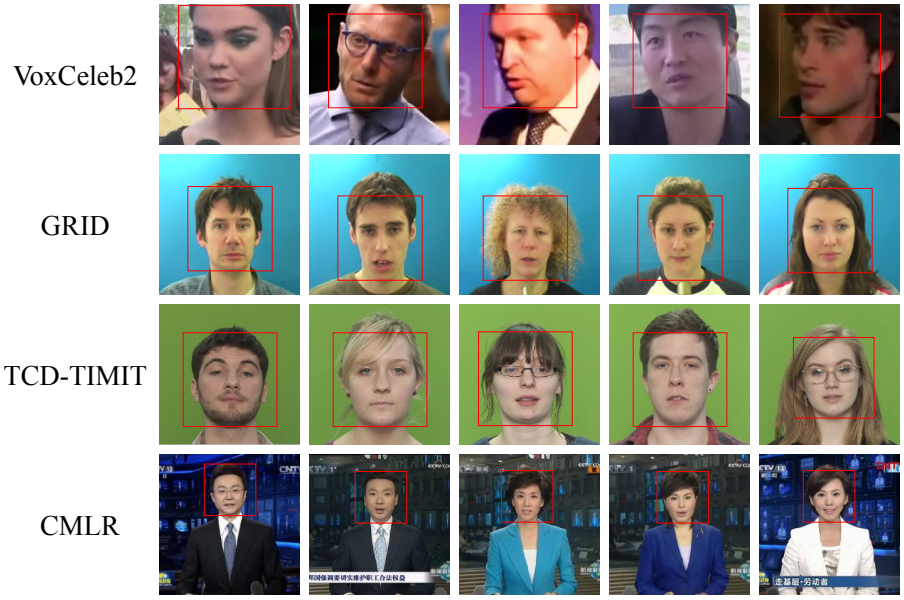


Figure 5.3: Random face samples from audio-visual corpora. Only the face region is cropped during training and test.

### 5.3.2 Evaluation Metrics

We evaluate the generated speech quality and intelligibility with perceptual evaluation of speech quality (PESQ) [159] and extended short-time objective intelligibility (ESTOI) [82] respectively. The speech-to-text results are measured with word error rate (WER) and character error rate (CER), the ratio of error terms, i.e., substitutions, deletions and insertions, to the total number of words/characters in the ground truth sequences.

Table 5.2: Overview of all corpora used in this chapter. Spk: Speakers. Utt: Utterances. Vocab: Vocabulary.

| Language | Dataset | #Spk. | #Utt. | #Vocab. | #hours | Usage | Modality |
|---|---|---|---|---|---|---|---|
| Multi-Language | VoxCeleb2 [35] | 6112 | 1.1M | - | 2442 | LipSound2 pre-training | Audio-Visual |
| English | GRID [41] | 51 | 33k | 51 | 27.5 | LipSound2 fine-tuning | Audio-Visual |
| | TCD-TIMIT [71] | 59 | 5.4k | 5.9k | 7 | | |
| | LJSpeech [81] | 1 | 13.1k | - | 24 | WaveGlow trainig | Audio |
| | LibriSpeech [138] | 2484 | 292.3k | - | 960 | Acoustic model pre-training | |
| Chinese | CMLR [220] | 11 | 102k | 3.5k | 87.7 | LipSound2 fine-tuning | Audio-Visual |
| | AISHELL-2 [48] | 1991 | - | - | 1000 | Acoustic model pre-training | Audio |

### 5.3.3 Training

We only describe the training settings of LipSound2 pre-training, LipSound2 fine-tuning and Jasper acoustic model fine-tuning. More details about Japser[1] pre-training acoustic model, KenLM[2] language model and WaveGlow[3] can be found on the open source websites.

#### Vision Stream

Face landmarks are detected using MTCNN [215] from all video frames and only the face area is cropped and reshaped to size of $112 \times 112$ as inputs. We also add one 'visual period' – an empty frame with all values of 255 – at the end of every visual stream to help the decoder stop decoding at the right time. A max decoder step threshold of 1000 is activated to terminate decoding when the decoder fails to capture the 'visual period'.

#### Audio Stream

We first divide the raw waveforms by the max value to normalize all audio to $[0, 1]$, then extract the magnitude using the short time Fourier transform (STFT) with 1024 frequency bins and a 64ms window size with 16ms stride. The Mel-scale spectrograms are obtained by applying an 80 channel mel filter bank to the magnitude, followed by dynamic range clipping with a minimum value of 1e-5 and log dynamic range compression.

[3]https://github.com/NVIDIA/waveglow

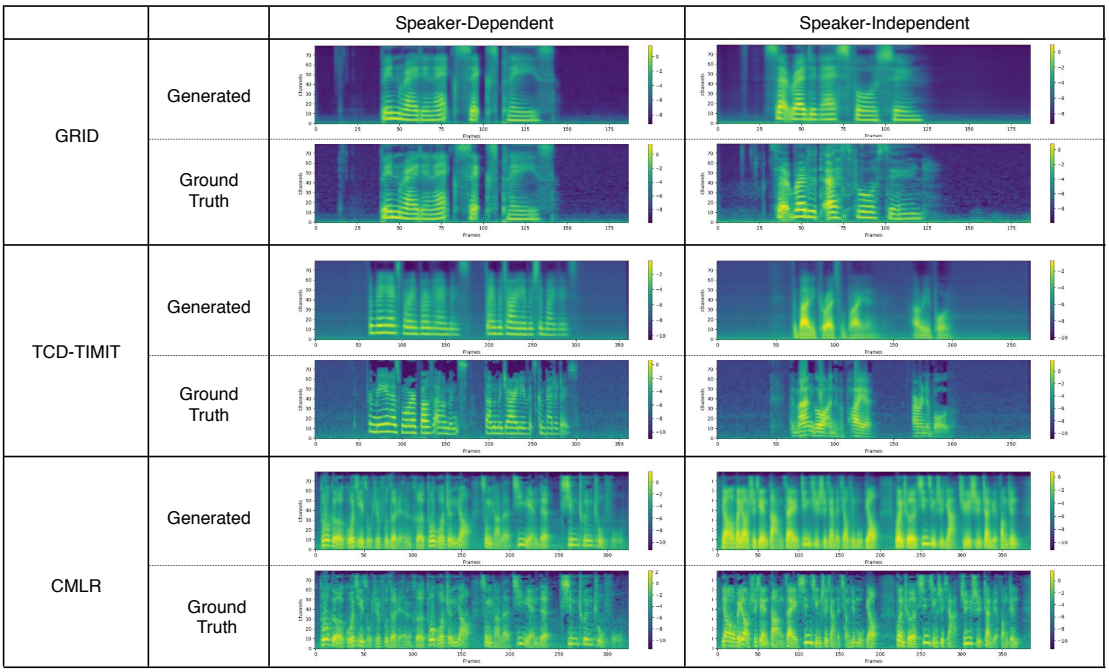


Figure 5.4: The comparison between generated Mel-spectrogram and ground truth in speaker-dependent and -independent settings for English and Chinese.

### LipSound2 Pre-training

Image horizontal flipping, gradient clipping with a threshold of 1.0, early stopping and scheduled sampling [17] are adopted to avoid overfitting. Linear and convolutional layers are initialized with Xavier [56] and tanh functions respectively. We use the cosine learning rate decay strategy with an initial value of 0.001. All experiments are conducted on 4 NVIDIA Quadro RTX 6000 GPUs with 24G memory.

### Fine-tuning

Pre-trained LipSound2 is fine-tuned on GRID, TCD-TIMIT and CMLR videos respectively to conduct speech reconstruction experiments. Afterwards, the produced speech for English (GRID and TCD-TIMIT) and Chinese (CMLR) is fine-tuned on the pre-trained English (LibriSpeech) and Chinese (AISHELL-2) acoustic models to perform lip reading tasks with a 10 times smaller learning rate.

## 5.4 Results and Discussion

### 5.4.1 Lip to Speech Reconstruction

#### Speaker-dependent Result

We report the generated speech results in two perspectives, i.e. speech quality (PESQ) and speech intelligibility (ESTOI). For a fair comparison, we keep the same settings as previous works. For speaker-dependent tasks, all datasets are randomly split into 90:5:5 for training, validation and test sets on GRID (Speaker $S1-S4$) and TCD-TIMIT (Lipspeaker $1-3$). Different from previous works that build one model for each individual speaker, we train only one model on all speakers to make it more convenient and closer to practical applications.

As shown in Table 5.3, our LipSound2 system which is firstly pre-trained on the VoxCeleb2 dataset, then fine-tuned on the specific dataset achieves highest scores on both PESQ and ESTOI, which reveals the effectiveness of our proposed method.

Table 5.3: Speaker-dependent speech reconstruction results on GRID and TCD-TIMIT datasets.

| | GRID | | TCD-TIMIT | |
|---|---|---|---|---|
| Model | ESTOI | PESQ | ESTOI | PESQ |
| Vid2Speech [51] | 0.335 | 1.734 | 0.298 | 1.136 |
| Lip2AudSpec [4] | 0.352 | 1.673 | 0.316 | 1.254 |
| Vougioukas et al. [188] | 0.361 | 1.684 | 0.321 | 1.218 |
| Ephrat et al. [49] | 0.376 | 1.825 | 0.310 | 1.231 |
| Lip2Wav [151] | 0.535 | 1.772 | 0.365 | 1.350 |
| vid2voc-M-VSR [128] | 0.455 | 1.900 | - | - |
| LipSound2 | **0.592** | **2.328** | **0.372** | **1.490** |

#### Speaker-independent Result

For speaker-independent cases, we follow the same setups for GRID [188] and TCD-TIMIT [71]. LipSound2 achieves the best results on both metrics on the GRID dataset. Moreover, by listening to the reconstructed audio, we find that our model is capable of producing similar voices as ground truth speakers, instead of

generating a weird voice or one of the voices in the training set as occurring in previous works. The model has implicitly learnt the mapping between voices and faces. We highly recommend readers to listen to the produced samples on our demo website[4].

Furthermore, we find substitution errors occurring on segment level (vowels and consonants) because the context information is still not sufficient to disambiguate the phonemes that share the same visible organs, like lips and tongue, but are different in the invisible ones.

To the best of our knowledge, we are the first to tackle the speaker-independent case on the TCD-TIMIT dataset, since TCD-TIMIT consists of limited samples (∼370) for each speaker but with large-scale vocabulary (∼5.9K), which makes the tasks on TCD-TIMIT quite challenging. The speaker-independent results reported in Table 5.4 show considerable performance, for example, the PESQ result is even better than some results reported on speaker-dependent settings (as shown in Table 5.3), which suggests that the large-scale self-supervised pre-training enables the model to successfully generalize to unseen speakers.

Table 5.4: Speaker-independent speech reconstruction results on GRID and TCD-TIMIT datasets.

| | GRID | | TCD-TIMIT | |
|---|---|---|---|---|
| Model | ESTOI | PESQ | ESTOI | PESQ |
| Vougioukas et al. [188] | 0.198 | 1.24 | - | - |
| vid2voc-M-VSR [128] | 0.227 | 1.23 | - | - |
| vid2voc-F-VSR [128] | 0.210 | 1.25 | - | - |
| LipSound2 | **0.363** | **1.72** | **0.30** | **1.31** |

### Speech Reconstruction for Chinese

To explore the effectiveness of our proposed architecture, we further perform speech reconstruction in Chinese. For the speaker-dependent case, we keep the same training and test splits used in CSSMCM [220] for lip reading; for the speaker-independent case, $S1$ (male) and $S6$ (female) are used for testing and the remaining speakers are used for training and validation.

In Table 5.5, only LipSound2 results are reported since we make a first attempt at tackling speech reconstruction in Chinese. After checking the generated audio

[4]https://leyuanqu.github.io/LipSound2/

samples, we find that, besides the confusion on segments, there are some tone errors. One of the reasons is that Chinese is a tonal language in which lexical tones play an important role for semantic discrimination. The fundamental frequency (F0) which is produced by the vibration of vocal cords is not visible in the input videos (face area), and it is reported that the visual features have a weak correlation to F0 [42]. Another reason is that the VoxCeleb2 dataset mainly consists of non-tonal languages, e.g. British English, American English and German, which makes the pre-training pay little attention to tone production.

Table 5.5: Speech reconstruction results for Chinese on CMLR datasets.

| | Speaker-dependent | | Speaker-independent | |
|---|---|---|---|---|
| Model | ESTOI | PESQ | ESTOI | PESQ |
| LipSound2 | **0.36** | **1.43** | **0.28** | **1.21** |

**Attention Alignment**

We compare the attention alignments learned by LipSound [154] which is only trained on the GRID dataset and LipSound2 (this chapter). As shown in Fig. 5.5, the LipSound attention weights are fuzzy at non-verbal areas and at short pauses between words, which may mislead the decoder into focusing on irrelevant encoder timesteps, whereas the attention weights learned by LipSound2 are intensive and more robust to silence or short pauses.

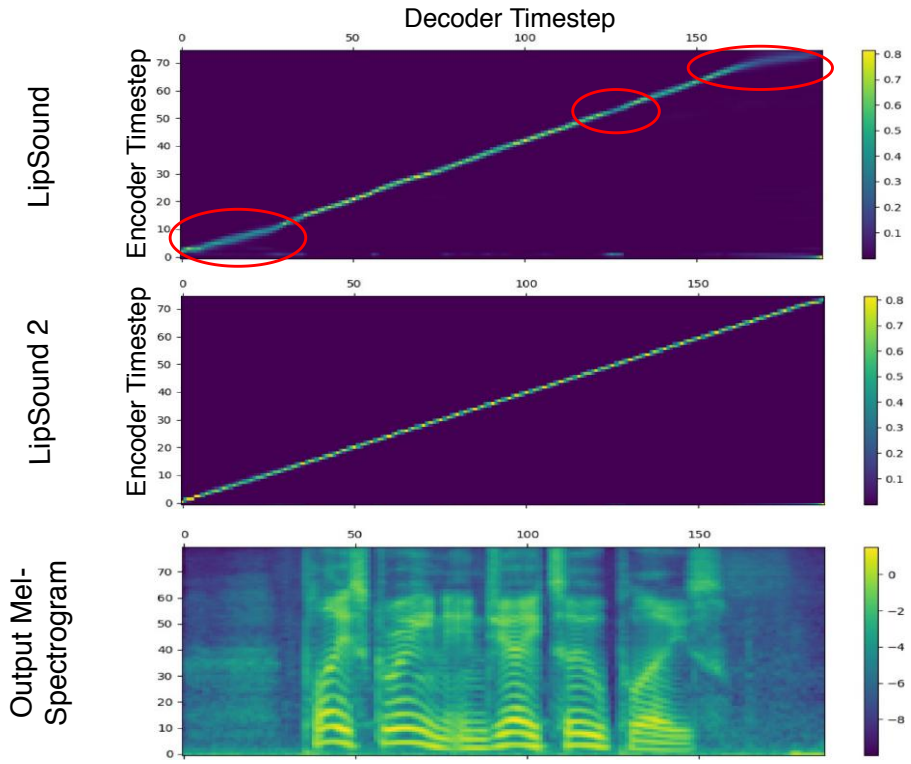


Figure 5.5: Attention alignment comparison on GRID dataset.

### 5.4.2 Lip Reading Results

Different from conventional methods which directly transform videos into text, we perform lip reading experiments in two steps, i.e. video-to-wav and wav-to-text.

#### Lip Reading Results for English

We follow the same splits as previous works for training and test on GRID [10] and TCD-TIMIT [179] datasets. The comparison with related results are listed in Table 5.6. We report the WER of GRID and TCD-TIMIT audio test sets on pre-trained acoustic models (Audio Gold Standard) and the results fine-tuned on the training audio samples (+Fine-Tuning), which is treated as the upper boundary of lip reading.

Our LipSound2 model achieves state-of-the-art performance on both GRID and TCD-TIMIT datasets. Fine-tuning the acoustic model pretrained on 960h LibriSpeech with generated audio can not only significantly boost the model performance but also accelerate training time.

Further improvement can be achieved when an external language model is integrated. The benefit from the language model on the GRID dataset is not as

Table 5.6: Lip reading results on GRID and TCD-TIMIT dataset on WER. Spk-Dep: Speaker-Dependent. Spk-Indep: Speaker-Independent. LM: Language Model.

| | GRID | | TCD-TIMIT | |
|---|---|---|---|---|
| Model | Spk-Dep | Spk-Indep | Spk-Dep | Spk-Indep |
| Audio Gold Standard | 22.36 | 21.88 | 15.86 | 15.21 |
| +Fine-tuning | 0.15 | 0.35 | 5.42 | 6.73 |
| LipNet [10] | 5.6 | 13.6 | - | - |
| LipNet+LM [10] | 4.8 | 11.4 | - | - |
| PCPG+LM [114] | - | 11.2 | - | - |
| TVSR-Net [209] | - | 9.1 | - | - |
| WAS [36] | 3.0 | - | - | - |
| LCANet[207] | 2.9 | - | - | - |
| DualLip [27] | 2.7 | - | - | - |
| LipSound [154] | 2.5 | - | - | - |
| CD-DNN [179] | - | - | 51.26 | 57.03 |
| MobiLipNetV2 [100] | - | - | - | 53.01 |
| LipSound2 | 1.9 | 7.3 | 41.37 | 46.29 |
| LipSound2 + LM | **1.5** | **6.4** | **39.77** | **43.53** |

much as on TCD-TIMIT, since the sentence structure in GRID is designed by an artificial grammar. The language model can only help to correct misspelled words but cannot contribute grammatically or semantically.

### Lip Reading Results for Chinese

We also explore lip reading performance in Chinese, as shown in Table 5.7. Audio Gold Standard is directly evaluating the CMLR test set on a pre-trained acoustic model trained on 1000h AISHELL2 dataset. After fine-tuning with CMLR training audio, we get 3.88% CER and 4.89% CER for speaker-dependent and -independent cases respectively.

In comparison to other work, our LipSound2 model achieves better results. CER further drops when decoding with an external language model. Besides, we build a new baseline for CMLR in speaker-independent settings.

Table 5.7: Lip reading results for Chinese on CMLR datasets. CER: character error rate.

| Model | Spk-dep | Spk-indep |
|---|---|---|
| Audio Gold Standard | 19.25 | 16.2 |
| +Fine-tuning | 3.88 | 4.89 |
| WAS [36] | 38.93 | - |
| CSSMCM [220] | 32.48 | - |
| LIBS [221] | 31.27 | - |
| LipSound2 | 25.03 | 36.56 |
| LipSound2 + LM | **22.93** | **33.44** |

## 5.5 Summary

In this chapter, we have proposed LipSound2 which directly predicts speech representations from raw pixels. We investigated the effectiveness of self-supervised pre-training for speech reconstruction on large-scale vocabulary datasets, particularly for speaker-independent settings. Moreover, state-of-the-art results are achieved by fine-tuning the produced audio on a well pretrained speech recognition model for both English and Chinese lip reading experiments, since our two-step method benefits not only from the large-scale crossmodal supervision which enables the model to learn more robust representations and more different content information, but also from the advanced speech recognition architecture (acoustic and language models) which is pre-trained on abundant labeled data.

# Chapter 6

# Multi-modal Target Speech Separation with Voice and Face References

## 6.1 Introduction

Speech separation aims to recover a clean speech signal from a mixed signal produced by multiple speakers simultaneously, e.g. in a cocktail party environment. Despite the significant progress on speech separation technologies over the past few years [76, 194, 50], the label permutation problem is still challenging for the speech signal processing community. The label permutation problem arises from label ambiguity — the arbitrary order of multi-output — which leads to an inconsistent gradient update and makes a neural network hard to converge during training. According to whether additional information is available, approaches for solving the problem can be mainly divided into two categories: blind speech separation and target speech separation. The blind speech separation task is to isolate a clean output for each individual source signal without any other information about the observed speech mixture, as shown in Figure 6.1 (a). To alleviate the permutation problem, deep clustering [76] and its variant, a deep attractor network [28], were proposed to disambiguate the label permutation. Permutation invariant training [211] was presented to predict the best label permutation, whereas the unknown number of sources and invalid outputs are still big challenges in this direction [117]. The model described in this Chapter has been published in Qu et al [155].

Different from blind speech separation, target speech separation only recovers the desired single signal guided by auxiliary information, e.g. source directions

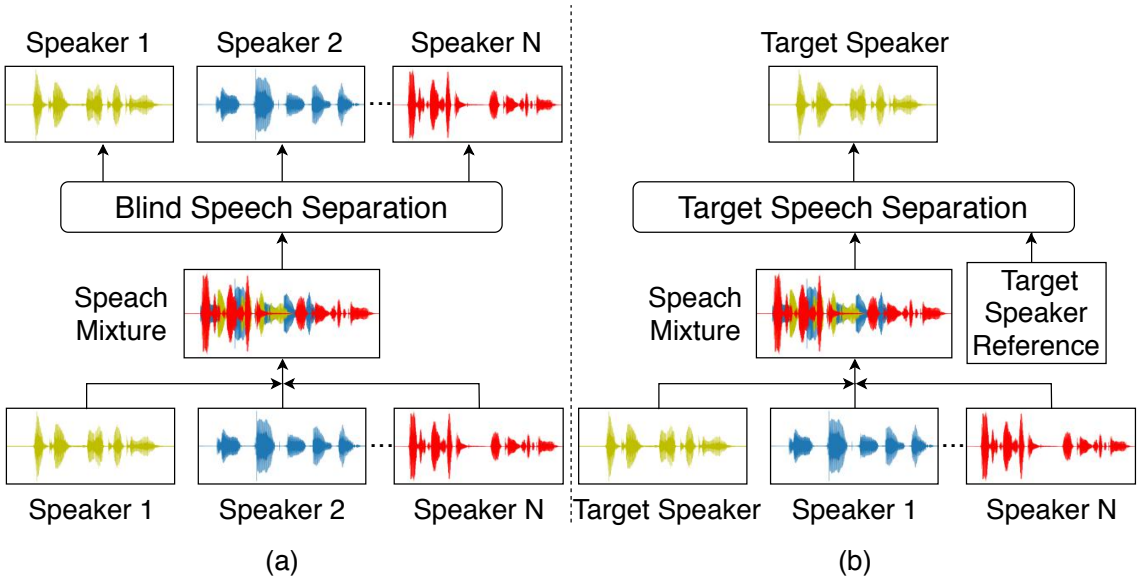


Figure 6.1: Comparison of (a) blind speech separation and (b) target speech separation.

or target speaker identity, as shown in Figure 6.1 (b). By leveraging the target speaker reference, target speech separation avoids the permutation problem and is independent of the number of source speakers, since there is only one output per time in this case.

More recently, multi-modal audio-visual approaches [50] have shown impressive results in target speech separation and attracted a lot of attention from the computer vision community, for instance, utilizing the lip movement sequences in videos to predict target time-frequency masks or directly generate the target waveform.

Inspired by VoiceFilter [193] which performed target speech separation with speaker voice embeddings and achieved good performance, in this chapter, we extend the audio-only VoiceFilter to the audio-visual domain and explore to what extent the visual modality (face embedding) can benefit target speech separation. Additionally, previous audio-visual methods strictly require simultaneous visual streams and highly depend on the visual temporal information. This is hard to meet in most real-world cases, because the speaker's mouth may be concealed by microphone [3] or be undetectable sometimes. Therefore, it is difficult to generalize the video-based methods to devices without cameras. To solve this problem, we propose to integrate the speaker face information into the system, which can be enrolled beforehand and easily applied to more challenging scenarios, for example, if an assistance robot works in public spaces with unknown people addressing it

for the first time, then their voice embedding is yet unavailable while their face image is available.

## 6.2 Model Architecture

As shown in Figure 6.2, our proposed model contains three neural networks: a pre-trained FaceNet for face embedding extraction, a pre-trained speaker verification net for voice embedding extraction, and a mask estimation net (the trainable modules) for target speaker mask prediction.

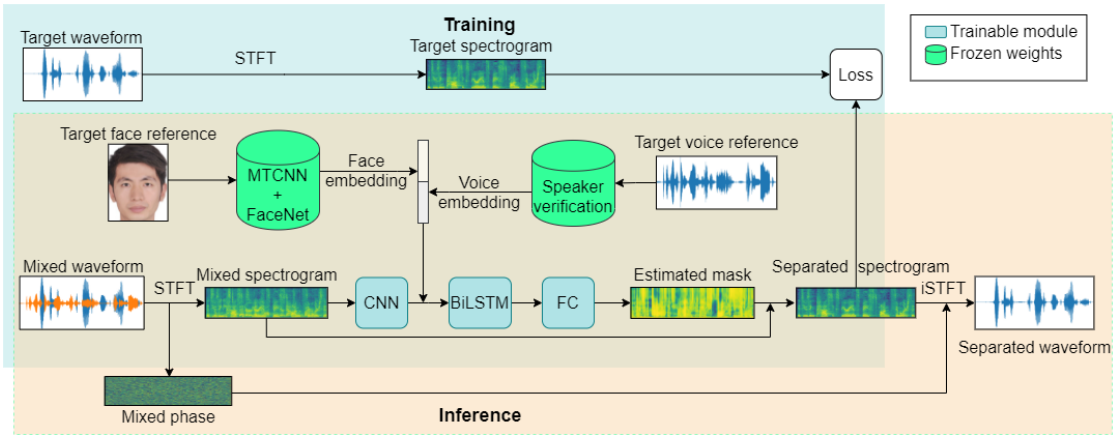


Figure 6.2: Overview of the proposed target speech separation architecture. The model receives inputs, i.e. the mixed spectrogram, the face embedding and/or the voice embedding to predict a target speaker time-frequency mask which is used to estimate the target spectrogram.

### 6.2.1 Face Embedding Net

The face embedding net is based on a Multi-task CNN (MTCNN) [215] and FaceNet [164] used in a sequence. Before feeding the original face images into FaceNet, an MTCNN is used for face detection, since the MTCNN performs better in some hard conditions, such as partial occlusion and silhouettes. We crop only the face region and reshape all faces to 160×160 size for face embedding extraction. FaceNet directly learns a unified embedding for different tasks, for example face recognition and face verification, and achieves good results on different benchmarks. In this chapter, we use FaceNet Inception-ResNet-v1 in Pytorch[1]. The model is pre-trained on the VGGFace2 dataset and achieves 99.65% accuracy on the evaluation set.

[1]https://github.com/timesler/facenet-pytorch

### 6.2.2 Voice Embedding Net

The voice embedding net is based on the model proposed by Wan *et al.* [189] for speaker verification, which consists of 3 LSTM layers with 768 nodes in each layer and one linear layer with 256-dimensional outputs. A generalized end-to-end loss was performed to cluster the utterances from the same class closer while increasing the distance between utterances from different classes during training. The pre-trained model[2] used in our thesis is trained on the VoxCeleb2 [35] dataset with thousands of speakers. The input spectrogram is extracted using the short time Fourier transform (STFT) with a 40ms hop length and a 80ms window size. The model achieves 7.4% equal error rate on the VoxCeleb1 test dataset (first 8 speakers).

### 6.2.3 Mask Estimation Net

The mask estimation net (the trainable modules in Figure 6.2) is to predict a target speaker mask in the time-frequency domain, which is heavily inspired by VoiceFilter [193] and the architecture proposed by Wilson *et al.* [203]. As shown in Table 6.1, the network begins with 7 Conv2D layers with different kernel sizes to capture the variations in time and frequency. Stacked dilated factors enable the network to have larger receptive fields. The output from the last CNN layer is concatenated with voice or/and face embeddings (repeated N times where N is the dimension of the spectrogram in time) as the input of the following bidirectional LSTMs layers. Two fully connected (FC) layers are used to map the high-dimensional outputs from LSTM to the dimension of spectrogram frequency. We use batch normalization and ReLU activation between each layer and a sigmoid function at the output layer. The separated spectrogram is obtained by multiplying the estimated mask and the mixed input. During inference, the separated waveform is reconstructed by the inverse STFT with the phase from the noisy mixture.

## 6.3 Experimental Setup

### 6.3.1 Dataset

We generate the training and test sets based on the lip reading sentences 3 (LRS3) dataset [2] which consists of thousands of speakers' videos from TED and TEDx.

[2]https://github.com/mindslab-ai/voicefilter

Table 6.1: Configuration of mask estimation network. Kernel is the kernel size in time and frequency. Dilation is the dilation factor in time and frequency.

| **Layer** | **Kernel** | | **Dilation** | | **Channels/Nodes** |
|---|---|---|---|---|---|
| | Time | Freq | Time | Freq | |
| CNN1 | 1 | 7 | 1 | 1 | 128 |
| CNN2 | 7 | 1 | 1 | 1 | 128 |
| CNN3 | 5 | 5 | 2 | 1 | 128 |
| CNN4 | 5 | 5 | 4 | 1 | 128 |
| CNN5 | 5 | 5 | 8 | 1 | 128 |
| CNN6 | 5 | 5 | 16 | 1 | 128 |
| CNN7 | 1 | 1 | 1 | 1 | 8 |
| BiLSTM1 | - | - | - | - | 400 |
| BiLSTM2 | - | - | - | - | 400 |
| FC1 | - | - | - | - | 601 |
| FC2 | - | - | - | - | 601 |

The dataset is transcribed on word level which will be used in our speech recognition experiments.

As shown in Figure 6.3, for the training set, we crop 3s clips from each video where the audio part is treated as the target speech and the visual part is used to get the speaker face from a random frame. To augment the face variants, 10 random faces are extracted from each visual part. The 10 faces are completely out of order and only one face is visible at a time during training. The mixed speech is simulated by directly adding the same length speech from a different random speaker to the target speech. The voice embedding is extracted from a different utterance by the same speaker. Finally, we get 200k samples for around 2k speakers.

For the test set, we use the same process but keep the utterance length in the LRS3 test set and discard the speakers who have only one utterance or the utterance length is less than 3s. Finally, we get 1171 utterances for 270 speakers. There is no speaker overlap between training and test sets.

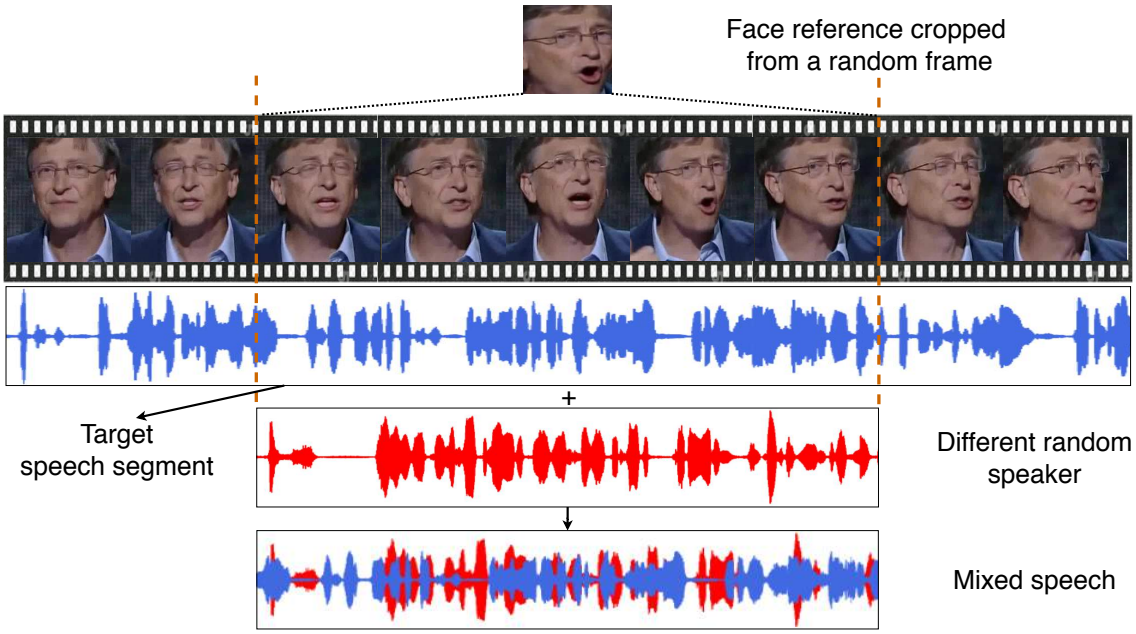


Figure 6.3: Dataset building.

### 6.3.2 Training

All experiments are conducted on a single NVIDIA Quadro RTX 6000 GPU with 24G memory. We use the Adam optimizer with an initial learning rate of 0.001 and anneal the learning rate with a value of 1.1 after every epoch. Subsequently, we extract 601-dimension Mel-spectrograms with a 25ms window size and a 10ms hop length from mixed speech as model input. Normalization is performed for each Mel-frequency bin with the mean and variance.

### 6.3.3 Evaluation Metrics

We evaluate the model performance with two metrics: source to distortion ratio (SDR) [186] and word error rate (WER). SDR[3] relates the estimated target signal to the noise terms and was found to negatively correlate with the amount of noise left in the separated audio signal [50]. We also evaluate the signal quality with WER by feeding the separated speech into a *Jasper* [110] speech recognition system which is trained on the 960h LibriSpeech dataset and achieves 3.61% WER on LibriSpeech dev-clean set. The evaluation is performed based on the OpenSeq2Seq[4] toolkit published by NVIDIA.

[3]http://craffel.github.io/mir_eval/
[4]https://nvidia.github.io/OpenSeq2Seq/html/speech-recognition.html

## 6.4 Results and Discussion

### 6.4.1 Results of Speech Separation

We visualize the voice and face embeddings from 14 random speakers in the training set. The face embeddings are 10 times more than voices since we randomly crop 10 face images for each mixed speech. As shown in Figure 6.4, the voice embedding points belonging to the same speakers tend to gather together and significantly far away from other classes. However, the face embedding points from the same speaker are dispersed and close to other classes. We found this is caused by different face angles since all videos are in the wild and the speaker may turn the head from left to right profile while talking.

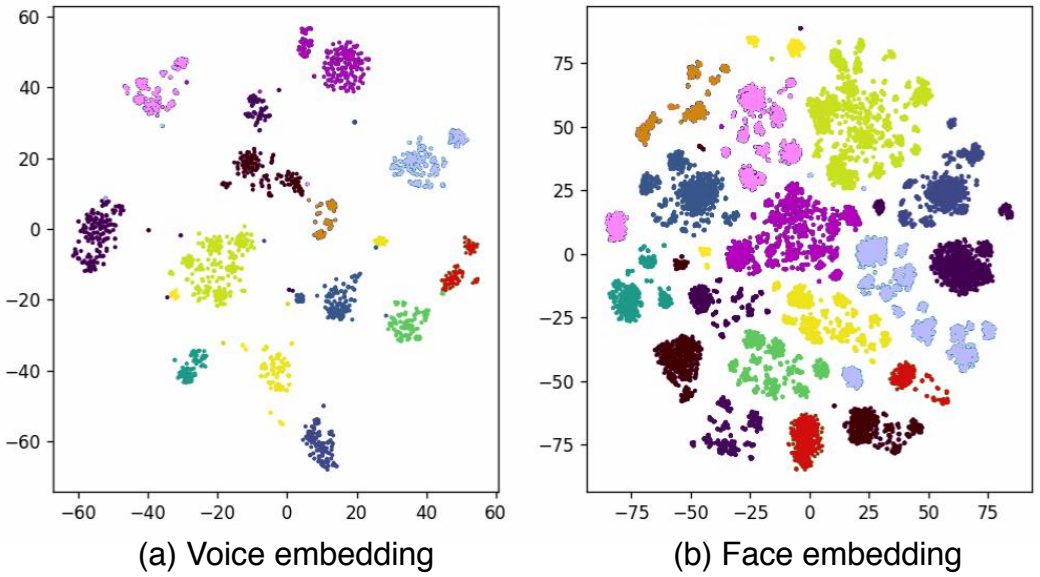


Figure 6.4: The visualization of (a) voice and (b) face embeddings for 14 randomly chosen speakers in training set with t-SNE. The face embedding points are relatively dispersed compared to the voice embeddings.

To investigate the effect of head poses on our experiments, we randomly extract 10 faces for each sample during inference. As shown in Table 6.2, the performance of using face embeddings fluctuate wildly according to different head poses (Std Dev: 0.32).

Compared to the result of only using voice embedding (10.32±0.11 dB), face information achieves competitive performance (9.23±0.32 dB). The quality of separated speech can be further improved by combining both face and voice references. After checking the output audio, we find that face and voice embeddings are com-

plementary in some cases — in other words, when two voices sound similar, the corresponding faces may be distinguishable, for example, with different skin colors.

Table 6.2: Source to distortion rate results for models using only-voice embedding, only-face embedding and both voice+face embeddings (higher is better).

| **Reference** | **SDR (dB)** |
|---|---|
| Voice | 10.32±0.11 |
| Face | 9.23±0.32 |
| Voice+Face | 10.65±0.28 |

### 6.4.2 Results of Speech Recognition

We test the speech recognition results by Jasper in three settings, clean speech input, mixed speech input and speech separated by our proposed model. The Jasper system achieves 11.8% WER on the clean inputs, whereas the performance dramatically drops down to 71.2% WER when using mixed speech input.

Table 6.3: Word error rate on Jasper speech recognition system.

| **Input Speech** | **Model** | **WER(%)** |
|---|---|---|
| Clean Speech | - | 11.83 |
| Mixed Speech | - | 71.22 |
| Separated Speech (Clean) | Voice | 13.46±0.08 |
| | Face | 15.31±0.19 |
| | Voice+Face | 13.36±0.12 |
| Separated Speech (Mixed) | Voice | 25.60±0.11 |
| | Face | 29.94±0.25 |
| | Voice+Face | 23.32±0.12 |

We investigate the separated speech inputs for ASR in two conditions. One is the Separated Speech (Clean) in which we test the performance of our proposed

model with clean speech input. A robust speech separation system should not only recover desirable output from a mixture, but also have good performance on clean speech input. Table 3 lists the similar results for voice (13.46±0.08% WER), face (15.31±0.19% WER) and voice+face (13.36±0.12% WER) versus clean input (11.83% WER). The other is the Separated Speech (Mixed) in which the ASR receives the separated speech from mixed signals. We can see, in Table 6.3, the ASR performance can be significantly improved by feeding enhanced speech compared to the 71.22% WER when directly using noisy speech as input. The speech separation system using voice embedding is superior to the one using face embedding. Combining both voice and face references achieves the lowest WER, which is consistent with the evaluation on SDR.

## 6.5 Summary

In this chapter, we propose a novel approach of integrating pre-enrolled face information into the target speech separation task. Our model avoids the speaker permutation problem and the problem of an unknown number of source speakers, which audio-only approaches suffer from. In addition, different from the conventional audio-visual speech separation methods which heavily rely on the temporal information from the visual sequences, our system can also be easily adapted to those devices without cameras or to scenarios where no simultaneous visual streams are available. The experimental results on speech separation and speech recognition reveal the effectiveness of face information and the complementarity to voice embeddings.

# Chapter 7

# Combining Articulatory Features with End-to-End Learning in Speech Recognition

## 7.1 Introduction

End-to-end learning has been successfully applied in many domains, such as handwriting recognition [106], neural machine translation [13], and so on. Furthermore, end-to-end models have become popular in automatic speech recognition (ASR) tasks. The conventional ASR pipeline consists of many different components: an acoustic model, a pronunciation model and a language model. These components are separate and require lots of human expertise, e.g. a handcrafted pronunciation dictionary and designed senone states for hidden Markov models (HMMs). Additionally, the training targets and alignment information needed for neural networks in a DNN-HMM paradigm can only be obtained from another GMM-HMMs (GMM is short for Gaussian Mixture Model) model which is trained beforehand. Such a pipeline requires not only multiple training stages but also different optimization functions [126].

To simplify this complex paradigm, end-to-end learning approaches [126, 61, 224, 34, 14, 24] have been proposed to replace hand-designed feature engineering and jointly learn all components in a single architecture. These approaches can be transformed into computational flow graphs which can be optimized by back-propagation in a simple end-to-end training process. End-to-end models are able to naturally handle sequences of arbitrary lengths and directly optimize the word error rate. However, it is challenging to integrate domain knowledge into these

models. Therefore, the goal of this study is to combine articulatory features into end-to-end learning.

Articulatory features (AFs), also known as phonological features, phonological attributes or distinctive phonetic features, are used to represent the movement of different articulators, such as lips and tongue, during speech production. AFs can be robustly estimated from speech by statistical classifiers, such as GMM and neural networks [91]. A series of studies have demonstrated that AFs can improve the performance of ASR systems by systematically accounting for coarticulation, speaking styles and other variability, especially in a noisy scenario [94]. Conventional methods to extract AFs from speech require precise boundary transcription. To get this boundary information, the usual practice is using forced alignments generated by a GMM-HMMs model [212], or labeling data manually at a frame level [163], which are complex and time-consuming.

Our hypothesis in this chapter is that AFs can provide useful and complementary representations that cannot be learned automatically by an end-to-end architecture. This thesis explores two approaches to integrate domain knowledge to improve end-to-end model performance. Our contribution is twofold: In the first step, we train a bank of AF extractors using connectionist temporal classification (CTC) in an end-to-end way, which does not require a precise phone or frame level boundary information. In the second step, we propose two approaches (fine-tuning networks and progressive networks) to integrate domain knowledge (articulatory features) into end-to-end learning in speech recognition tasks. The model described in this Chapter has been published in Qu et al [153].

## 7.2 Model Architecture

In this section, we present the details of AF extractors, fine-tuning networks and progressive networks.

### 7.2.1 AF Extractor

Figure 7.1 shows the flow diagram to get AF-level transcriptions. First, we split words into phonemes according to the CMU dict[1]. Then, we generate AFs transcriptions according to the mapping [212] (see Table A.1 in the Appendix). The AF-level transcriptions will be used as training targets to build the AF extractors.

[1]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

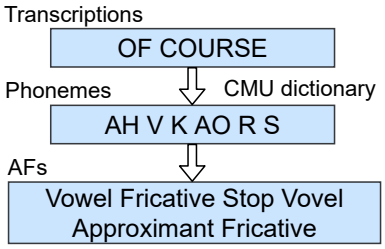


Figure 7.1: Flowchart to convert word level transcriptions of the phrase "of course" to AF labels.

Eight AF extractors were built: place, manner, anterior, back, continuant, round, tense and voiced. The AF extractor architecture is shown in Figure 7.2 (a), which begins with two layers of 2D convolutions, followed by five layers of gated recurrent units (GRU), and the output layer is fully connected. We train each extractor with the CTC and additional two symbols (blank and space). For example, for 'voiced', the target labels are {*voiced, other, space, blank*}.

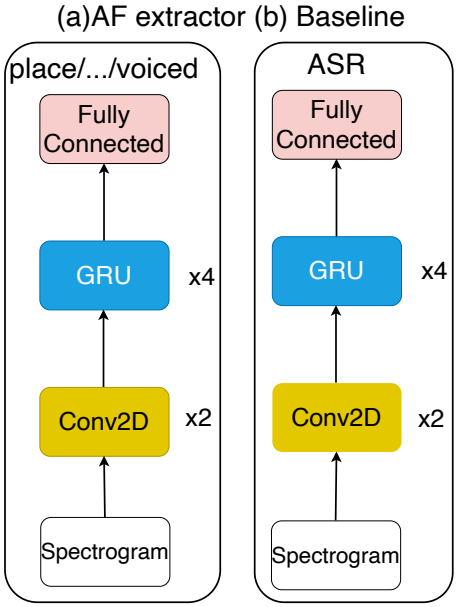


Figure 7.2: Illustration of (a) AF extractor, (b) ASR baseline system. The ASR baseline system is based on Deep Speech 2 [6].

### 7.2.2 Fine-tuning Network

Fine-tuning is a process to transfer what a neural network learned on a given task to a second task. AF extractors learned in a first task can be treated as a fixed

front-end to transform spectrograms to AFs. Hidden layer outputs from different AF extractors will be combined, then fed into another neural network for the second task (ASR). Figure 7.3 (a) and (b) show the fine-tuning networks used in this study. The details of AF extractors (place, manner, anterior, back, continuant, round, tense and voiced) are shown in Figure 7.2 (a). We concatenate the fourth or fifth GRU layer output of all extractors as a vector, namely fine-tuning network 1 (Figure 7.3 (a)) and fine-tuning network 2 (Figure 7.3 (b)) respectively, and feed it into a 5 bidirectional GRU-layer neural network for the ASR task.

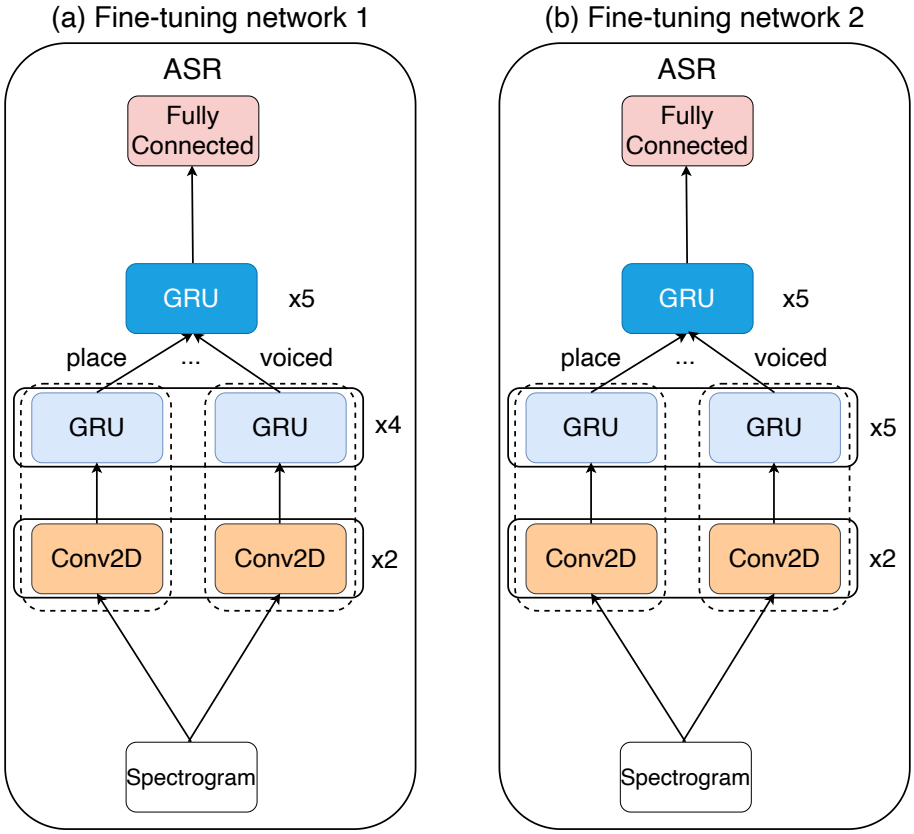


Figure 7.3: Illustration of (a) fine-tuning network 1 and (b) fine-tuning network 2. Note: frozen (dotted line) without backpropagation and weight updating.

### 7.2.3 Progressive Network

The progressive networks with lateral connections from previous tasks can accelerate learning speed and avoid forgetting [162]. They not only learn relevant features but also acquire different representations from previous learned tasks, which may be irrelevant to the target task. The scheme of progressive networks are shown in Figure 7.4. There are no connections between the AF extractors and they are

trained in parallel and independently, then linearly combined. The source task is AF extraction from speech signals and the target task is speech recognition. We use the following formula to compute outputs of layer $i$ in ASR tasks:

$$h_i = W_i(h_{i-1} + \sum_{j=1}^{8} k_{i-1}^{j}) \tag{7.1}$$

where $h_i$ is the output of layer of the ASR system, $k_{i-1}^{j}$ is the output of layer $i$ of AF extractor $j$, $W_i \in R_j^{i-1}$ is the weight matrix of layer $i$ of ASR systems, with $n_i$ the number of units at layer $i$. Layer $h_i$ receives input from both $h_{i-1}$ and $k_{i-1}^{j}$ via Eq. (7.1).

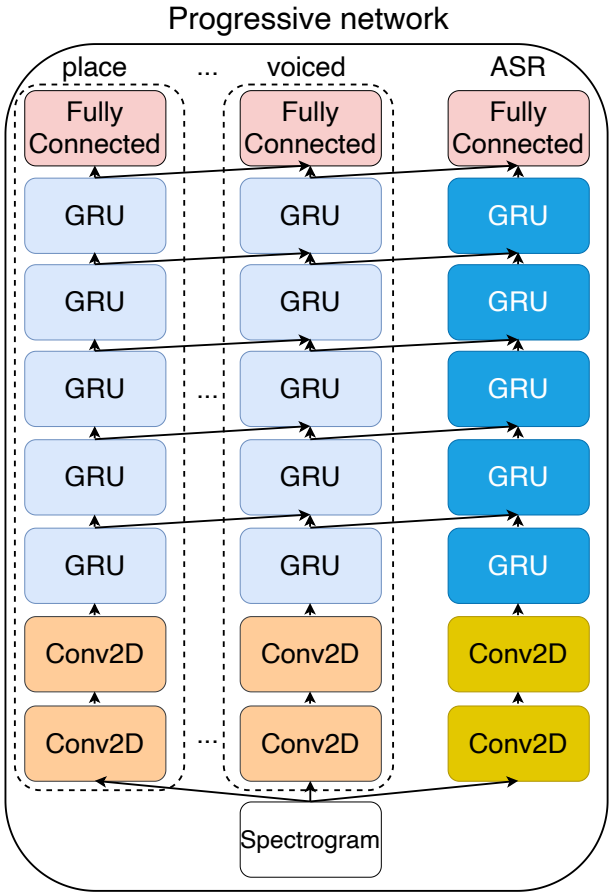


Figure 7.4: Illustration of progressive network. Note: frozen (dotted line) without backpropagation and weight updating.

## 7.3 Experiments

In this section, we present the dataset and the experimental setup.

### 7.3.1 Data

We use the Wall Street Journal (WSJ) [143] speech corpus both for AF and ASR experiments. The training set is the 81 hours 'train-si284' with about 37K sentences. We use the 'dev93' development set for validation and hyperparameter optimization and report the final performance on the 'eval92' test set.

### 7.3.2 Training

The baseline ASR system (shown in Figure 7.2 (b)) used in this chapter is similar to the Deep Speech 2 system [6]. The first two layers of all architectures are 2D (frequency and time domains) convolutions. The convolution layers not only reduce temporal variability in the time domain but also normalize speaker variance in the frequency domain. These are followed by GRU layers. It has been shown that GRU cells achieve comparable performance to long short-term memory (LSTM) but GRU cells are faster and easier to train [83]. Finally, we pass the output from the GRU cells to a fully connected layer.

The input features for all models are spectrograms derived from the raw audio files, with a 20ms window size and a 10ms window stride. All neural networks are trained with the CTC, using the stochastic gradient descent optimization strategy along with a mini-batch of 20 utterances per batch. We use 40 epochs and pick the model that performs best on the development set to evaluate the test set. Learning rates are chosen from [1e-4, 6e-4], and a learning rate annealing algorithm is used by the value of 1.1 after each epoch. The momentum is 0.9. Batch normalization is used to optimize models and accelerate training on hidden layers. All architectures described in this chapter do not use language models and add 'space' to segment outputs into words. The output alphabet for ASR experiments consists of 29 classes (a, b, c, . . . , z, space, apostrophe, blank). Once all AF extractors have been built, we freeze all extractor weights during ASR training. All models are trained on the corpus described in 4.1.

## 7.4 Results and Discussion

In this section, we present the performance of AF extractors and ASR systems using fine-tuning networks and progressive network. Table 7.1 shows the error rate of different AF extractors trained on the 81 hours 'train-si284' training set. All error rates are less than 10%, from which we conclude that articulatory features can be robustly detected from speech signals using the CTC loss function without

requiring boundary alignment information.

Table 7.1: Results of articulatory feature extractors at a phoneme level.

| | Articulatory Features | | Error Rate (%) |
|---|---|---|---|
| Place | Vowel<br>Fricative<br>Nasal | Stop<br>Approximant | 9.4 |
| Manner | Coronal<br>High<br>Dental<br>Glottal<br>Labial | Low<br>Mid<br>Retroflex<br>Velar | 8.6 |
| Others | Anterior | | 5.2 |
| | Back | | 9.2 |
| | Continuant | | 4.0 |
| | Round | | 9.1 |
| | Tense | | 8.7 |
| | Voiced | | 4.0 |

Table 7.2 lists the results from our ASR experiments and some results as reported in previous approaches using the CTC loss function on the WSJ benchmark. The fine-tuning network 1 (using 4-layer GRU from AF extractors) achieves a 33.2% WER which is worse than the baseline model (32.4%). However, when concatenating 5 layers of output from all AF extractors, the fine-tuning network 2 performs better than both the fine-tuning network 1 and the baseline system. We hypothesize that the deeper fine-tuning network 2 can capture more invariant and effective articulatory representation than the architecture with shallow layers.

The progressive network perform best in all our approaches achieving 28.6% WER. The progressive network can avoid forgetting and provide some complementary articulatory representations which can be learned by end-to-end architectures.

To examine the approaches we proposed and make a fair comparison, we cite some previous approaches which use CTC and an end-to-end architecture and only compare the ASR performance without additional language models. Compared to prior approaches, the final performance of our progressive network (28.6%) is

Table 7.2: Word error rate (WER) on the Wall Street Journal Corpus "eval92 20k" evaluation set. All models are trained with CTC loss function. No language models are used but the CTC-lexicon model [126] uses a lexicon.

| Model | WER(%) |
|---|---|
| RNN-CTC [61] | 30.1 |
| BDRNN-CTC [69] | 35.8 |
| CTC-lexicon [126] | 26.9 |
| Baseline | 32.4 |
| Fine-tuning network 1 | 33.2 |
| Fine-tuning network 2 | 31.6 |
| Progressive network | 28.6 |

better than the bidirectional RNN model [6] (35.8%) and the RNN-CTC approach (30.1%). It is not as good as the CTC lexicon system [126] (26.9%) which uses a lexicon in decoding and the lexicon helps to correct the output to correctly spelled words but we do not.

## 7.5 Summary

In this chapter, we have presented two approaches to combine domain knowledge, i.e. AFs, into end-to-end learning. First, fine-tuning neural networks are proposed to concatenate hidden layer outputs of AF extractors as inputs to another RNN for ASR. Second, a progressive neural network with lateral connections from AF extractors is proposed to integrate articulatory knowledge into an end-to-end architecture. Results show that both approaches can effectively incorporate articulatory information into end-to-end learning. Furthermore, the progressive neural network brings a significant improvement compared to the baseline system and previous works. In the following chapter, we would further explore how to leverage domain knowledge to recognize out-of-vocabulary words.

# Chapter 8

# Paying More Attention to Unseen Words: New Vocabulary Acquisition for End-to-End Speech Recognition

## 8.1 Introduction

Recently, end-to-end ASR models have been receiving a lot of attention and achieving impressive performance [61, 59, 13]. It significantly simplifies the training process to be able to directly map acoustic inputs to characters or words. Additionally, limited domain-specific knowledge is required, which dramatically boosts model development and deployment. However, the end-to-end models are heavily data-hungry and perform poorly on words out-of-vocabulary (OOV) or rarely existing in training data, for example, trending words and new named entities.

Current approaches for solving the OOV problem are mainly conducted on language model (ML) or post-processing, for instance, user-dependent language model [21, 121], LM rescoring [65] and finite-state transducer lattice extension [219], since it is expensive to collect labeled OOV speech data for ASR model training. However, the post-processing techniques only obtain limited improvement and it is hard to tackle the root causes acoustically.

Alternatively, fine-tuning ASR models with synthetic audio has shown promising performance on OOV recognition [222], which leverages the advanced text-to-speech (TTS) systems to generate audio-text pairs required by ASR model training. In this chapter, we take this method a step further and propose loss rescaling

to encourage models to pay more attention to unknown words. Instead of just fine-tuning ASR models in which all words are treated equally, enlarging the loss of utterance containing OOV words (sentence level) or increasing the gradient of unseen words (word level) can efficiently incline the model to update the weights related to OOV words. We randomly choose 5 trending words and 5 new named entities from the new words recorded by Cambridge Dictionary recent few years. Then, we crawl texts including the new words from the Internet and synthesize audio with TTS systems. The experimental results of fine-tuning audio-text pairs on a hybrid CTC/attention ASR model show a significant improvement on recall. When combining elastic weight consolidation with word level loss rescaling, we achieve 76.4% recall on the OOV test but with only 2.2% and 2.6% relative WER increase on the LibriSpeech test-clean and test-other respectively. The goal of this chapter is to improve the recognition of OOV words and keep the accuracy on non-OOV words.

## 8.2 Methodology

### 8.2.1 Loss Rescaling at Sentence Level

During training, CTC returns one loss per utterance and the mean of all utterance losses in the same mini-batch would be used for back-propagation. We observe that the utterance losses in one mini-batch are evenly distributed, as shown in Figure 8.1 (a). Consequently, the model equally pays attention to each utterance or word, which leads to the final model performance heavily relying on the frequency of words in training sets.

Sometimes, the model attention is even attracted immoderately by words not concerned. To emphasize the utterance containing OOV words, we rescale the utterance loss by multiplying a hyper-parameter $\mu$ in Eq. (8.1), where $x$ are acoustic inputs, $y$ are the target text references, and $O$ is the set of OOV words.

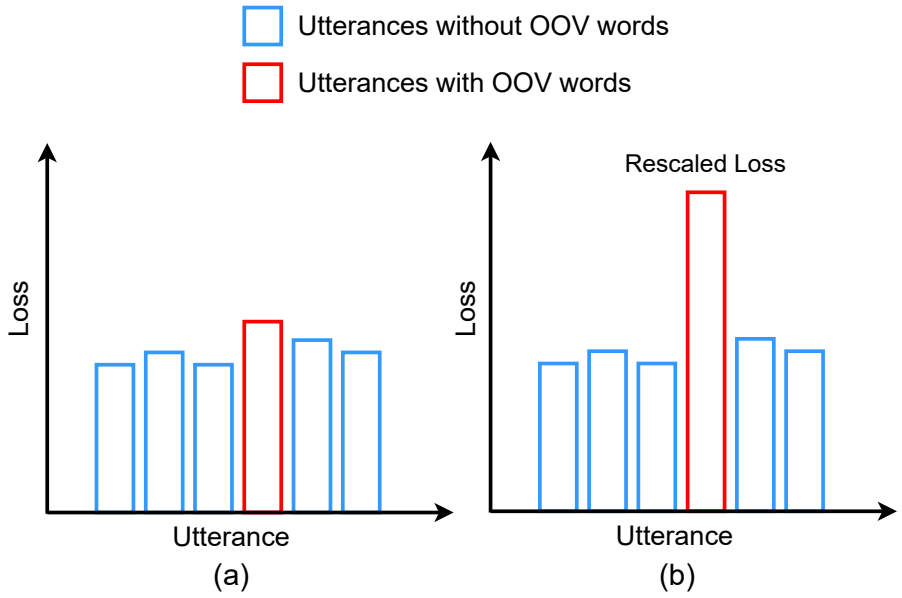


Figure 8.1: (a) Utterance loss distribution in one mini-batch. (b) Utterance loss distribution after loss rescaling.

$$\mathcal{L}_{sentence}(x,y) = \begin{cases} \mathcal{L}_{CTC}(x,y), & \text{if o not in y, } \forall o \in O \\ \mu\mathcal{L}_{CTC}(x,y), & \text{if o in y, } \forall o \in O \end{cases} \tag{8.1}$$

### 8.2.2 Loss Rescaling at Word Level

Given an input acoustic vector $\mathbf{x} = (x_0, \cdots, x_T)$, and a target label sequence $\mathbf{y} = (y_0, \cdots, y_U)$, the CTC loss aims to maximize the log probability in Eq. (8.2), where $\widetilde{\mathbf{y}}$ is the extended label sequence of $\mathbf{y}$ by inserting blank labels $\phi$ at the beginning and end of $\mathbf{y}$ and between every two label tokens, $\widetilde{\mathcal{Y}} = \mathcal{Y} \cup \phi$ and $\mathbf{a}$ is a possible token path output from neural networks.

$$\mathcal{L}_{CTC} = -logP(\widetilde{\mathbf{y}}|\mathbf{x}) = -log \sum_{\mathbf{a} \in \mathcal{F}^{-1}(\widetilde{\mathbf{y}})} P(\mathbf{a}|\mathbf{x}) \tag{8.2}$$

where $\mathcal{F} : \widetilde{\mathbf{y}} \rightarrow \mathbf{y}$ is the function that removes blank tokens and merges repeat labels to map the extended label sequence $\widetilde{\mathbf{y}}$ back to true label sequence $\mathbf{y}$.

We denote $\widehat{y}(t,u)$ and $b(t,u)$ as the probability of label and blank at node $(t,u)$ respectively. According to the definition of CTC [58], for any label tokens (the black nodes in Figure 8.2), for example the yellow node$(t_6, u_7)$, three red arrows can reach to the node. So the forward variable $\alpha(t,u)$ can be calculated recursively by adding the three possibilities as shown in Eq. (8.3). Similarly, three blue arrows

emit from node$(t_6, u_7)$ and the backword variable $\beta(t, u)$ can be represented by Eq. (8.4).

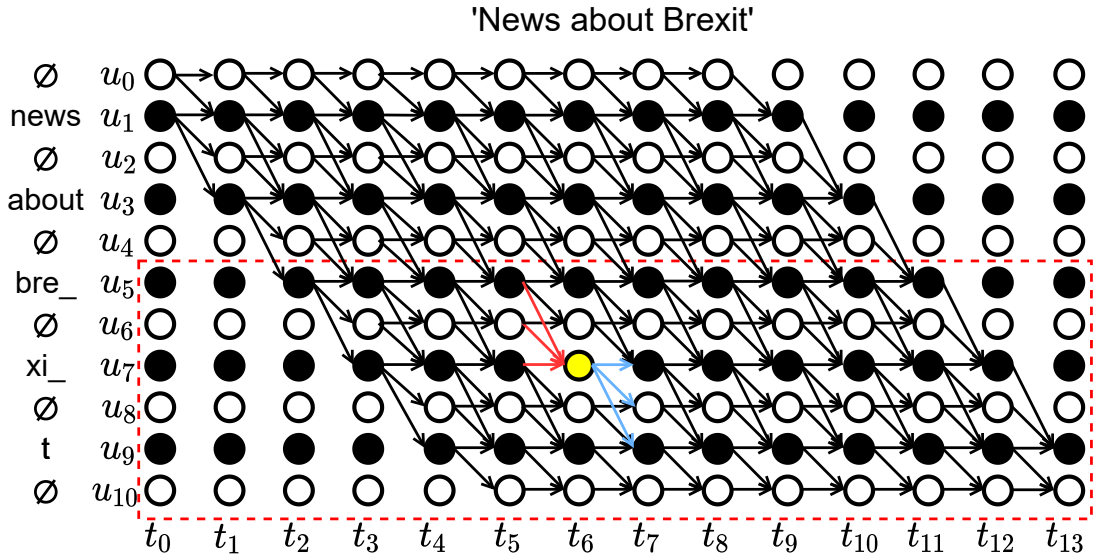


Figure 8.2: Illustration of CTC decoding lattices for the example sentence of 'News about Brexit', where the modeling unit is subword.

$$\alpha(t,u) = \widehat{y}(t-1,u-2)\alpha(t-1,u-2) + b(t-1,u-1)\alpha(t-1,u-1) + \widehat{y}(t-1,u)\alpha(t-1,u) \tag{8.3}$$

$$\beta(t,u) = \widehat{y}(t,u)\beta(t+1,u) + b(t,u)\beta(t+1,u+1) + \widehat{y}(t,u)\beta(t+1,u+2) \tag{8.4}$$

$P(\mathcal{A}_{t,u}|\mathbf{x})$, the probability of any candidate paths passing through node $(u,t)$ conditioning on the input sequence $\mathbf{x}$, can be obtained by multiplying forward (Eq. (8.3)) and backward probabilities (Eq.(8.4)).

$$\begin{aligned} P(\mathcal{A}_{t,u}|\mathbf{x}) &= \alpha(t,u)\beta(t,u) \\ &= \alpha(t,u)\widehat{y}(t,u)(\beta(t+1,u) + \beta(t+1,u+2)) \\ &+ \alpha(t,u)b(t,u)\beta(t+1,u+1) \end{aligned} \tag{8.5}$$

Thereby, the gradient of CTC loss function $\mathcal{L}$ *w.r.t* $\widehat{y}(t,u)$ and $b(t,u)$ can be estimated by Eq. (8.6) and Eq. (8.7) respectively.

$$\frac{\partial \mathcal{L}}{\partial \widehat{y}(t,u)} \propto \alpha(t,u)(\beta(t+1,u) + \beta(t+1,u+2)) \tag{8.6}$$

$$\frac{\partial \mathcal{L}}{\partial b(t,u)} \propto \alpha(t,u)\beta(t+1,u+1) \tag{8.7}$$

The CTC function treats all nodes equally and aims to minimize the global loss, which makes models hardly focus on local connections. To guide models paying more attention on the OOV words, we emphasize the OOV words (the nodes in the dotted box in Figure 8.2) by rescaling the probabilities of OOV nodes in candidate alignments. Thus, the rescaled probability of all alignments passing through OOV nodes can be presented as follows:

$$\widetilde{P}(\mathcal{A}_{t,u}|\mathbf{x}) = \begin{cases} \mu P(\mathcal{A}_{t,u}|\mathbf{x}), & \text{if } u \in O \\ P(\mathcal{A}_{t,u}|\mathbf{x}), & \text{otherwise} \end{cases} \tag{8.8}$$

The regularized loss function is shown in Eq. (8.9):

$$\widetilde{\mathcal{L}}_{CTC} = -log \sum_{\mathbf{a} \in \mathcal{F}^{-1}(\widetilde{\mathbf{y}})} \widetilde{P}(\mathcal{A}_{t,u}|\mathbf{x}) \tag{8.9}$$

We implement our approach by multiplying the gradients of OOV nodes on the candidate path with $\mu$, as shown in the following equations:

$$\frac{\partial \widetilde{\mathcal{L}}}{\partial \widehat{y}(t,u)} = \begin{cases} \mu \frac{\partial \mathcal{L}}{\partial \widehat{y}(t,u)}, & \text{if } u \in O \\ \frac{\partial \mathcal{L}}{\partial \widehat{y}(t,u)}, & \text{otherwise} \end{cases} \tag{8.10}$$

$$\frac{\partial \widetilde{\mathcal{L}}}{\partial b(t,u)} = \begin{cases} \mu \frac{\partial \mathcal{L}}{\partial b(t,u)}, & \text{if } u \in O \\ \frac{\partial \mathcal{L}}{\partial b(t,u)}, & \text{otherwise} \end{cases} \tag{8.11}$$

where $O$ is the set of OOV words tokenized into subwords.

### 8.2.3 Overcoming Catastrophic Forgetting

Directly fine-tuning models on a dataset obeying a different distribution from the original training set may lead to catastrophic forgetting. The updated model overfits the new dataset but forgets the knowledge learned on the original one. To overcome models suffering catastrophic forgetting, we adapt two approaches during fine-tuning. The first one is mixing partial original audio from LibriSpeech used for baseline model training with the synthetic speech, since adding the data obeying the same distribution with the training set can efficiently mitigate the overfitting problem. We explore the effect of different mixing ratios and present the results in Section 8.4. The other approach is constraining model parameters from updating

during fine-tuning with L2 or elastic weight consolidation (EWC) [95] and we will introduce the details in the following sections.

**L2 Regularization**

The L2 regularization loss $\mathcal{L}_{L2}(\theta)$ is shown in Eq. (8.12), where $\mathcal{L}_{CTC}(\theta)$ is the original CTC loss or rescaled loss in Eq. (8.1) and Eq. (8.9). $\theta_i$ is the $i$th parameter of the ASR model to be updated during fine-tuning, and $\theta_i^{'}$ is the $i$th parameter in the baseline model which is invariable and saved locally. $\lambda$ is the coefficient to balance the scale of two parts. L2 loss takes the difference between the fine-tuned model and the old model into account to ensure the updated model will not stray away too much from the baseline.

$$\mathcal{L}_{L2}(\theta) = \mathcal{L}_{CTC}(\theta) + \frac{\lambda}{2}\sum_i(\theta_i - \theta_i^{'})^2 \tag{8.12}$$

**Elastic Weight Consolidation**

Different from L2 loss that always refers to a fixed standard and treats all parameters equally, as shown in Eq. (8.13), the EWC loss uses the diagonal of Fisher information matrix $F$ to dynamically weigh the importance of each model parameter for the source task. During fine-tuning, the parameters important for source task remain unchanged to avoid knowledge forgetting whereas the less important ones are encouraged to update to fit on the new task. In this chapter, the diagonal of the Fisher information matrix is estimated on the LibriSpeech 960h training set.

$$\mathcal{L}_{EWC}(\theta) = \mathcal{L}_{CTC}(\theta) + \frac{\lambda}{2}\sum_i F_i \cdot (\theta_i - \theta_i^{'})^2 \tag{8.13}$$

## 8.3 Experiments

### 8.3.1 ASR Model Architecture

The end-to-end ASR model used in our experiments is the two-pass hybrid CTC/attention architecture, U2 [214], as shown in Figure 8.3. The shared encoder converts acoustic features $x$ into a latent vector $\mathbf{h}^{enc}$, then the CTC decoder transforms the latent vector into character/word probability $P(y_t|x_t)$ with the same length as input frames. In the meanwhile, the attention decoder generates one character/word probability $P(y_u|y_{u-1}, \cdots, y_0, x)$ per time step by conditioning on the attention content vector $\mathbf{c}_u$ and the decoder output from last step $y_{u-1}$. During training, the

sum of CTC loss and attention loss is used to do backpropagation, while during inference, the n-best hypotheses produced by the CTC decoder are rescored by the attention decoder to obtain better performance. The candidate with the highest score will be the final output.

We use the U2 model published in WeNet toolkit [213] which unifies streaming and non-streaming ASR models into one architecture by proposing the dynamic chunk-based attention. The encoder consists of 12 conformer blocks, with 4 multi-head attention, 2048 linear units, swish activation, a positional dropout rate of 0.1 and Conv2D kernel size of 31 for each block. The attention decoder contains 6 transformer blocks and the CTC decoder is composed of 1 linear layer and 1 log softmax function. The U2 model is pre-trained on the 960h LibriSpeech corpus and achieves 3.18% WER and 8.72% WER on test-clean and test-other respectively when rescoring with attention decoder. The pre-trained weights are available on website[1], which will be the baseline model for all our experiments.

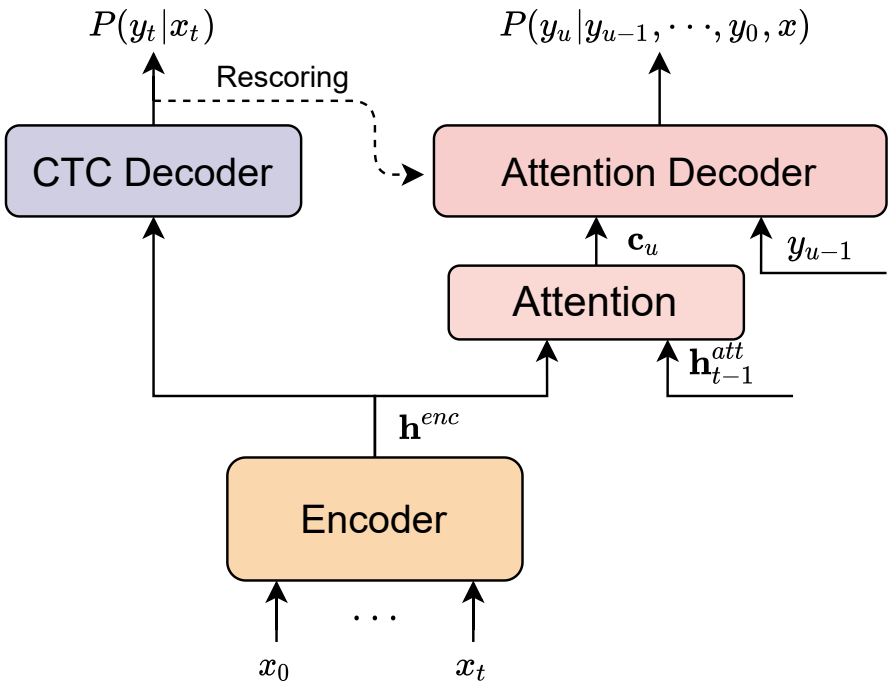


Figure 8.3: Two-pass hybrid CTC/attention ASR architecture.

[1]http://mobvoi-speech-public.ufile.ucloud.cn/public/wenet/librispeech/20210216_conformer_exp.tar.gz

Table 8.1: New words and crawled sentence examples.

| | **New words** | **Examples** |
|---|---|---|
| New trending words | Coronasomnia | Coronasomnia is the term used for sleep problems related to the pandemic. |
| | Maskne | Numerous factors may lead to the development of maskne. |
| | Brexit | Free trade agreements like the Brexit deal often include level playing field measures. |
| | Rollable | LG Rollable would feature a flexible OLED display from Chinese firm BOE. |
| | Blursday | I'll also suggest a mindful reframe of Blursday. |
| New named entites | Tiktok | The art of the TikTok comedy sketch is unique. |
| | Instagram | Instagram is also developing other tools aimed at protecting teen users. |
| | Spotify | Also make sure you are also running the latest version of Spotify on your iPhone. |
| | Duolingo | We hope that such people will do the Duolingo course out of curiosity. |
| | Paypal | PayPal Continues to Stand for Equality and Democracy. |

### 8.3.2 Text Crawling

We simulate the 10 OOV words dataset to evaluate our proposed approach since there is no standard OOV corpora published by the community. To ensure the words are definitely not in the LibriSpeech vocabulary, we randomly choose 5 trending words and 5 new named entities from new words recorded by Cambridge Dictionary[2] in the recent few years. We crawl 500 sentences for each new word with Scrapy[3] from the Internet, then remove from the sentences more than 50 words in case of running out of memory. Table 8.1 lists some trending words and new named entities used in this chapter. More examples can be found in Appendix E.

[2]https://dictionaryblog.cambridge.org/category/new-words/page/
[3]https://scrapy.org/

### 8.3.3 Speech Synthesis

We split the 500 sentences of each new word into training, validation, and test set with a ratio of 8:1:1. Instead of applying a multi-speaker TTS system which is more convenient and is able to produce more kinds of voices, we use some commercial APIs, i.e. Baidu TTS API[4], Google TTS API[5], iFLYTEK TTS API[6], Tencent TTS API[7] and Alibaba TTS API[8] to synthesize audio. In contrast to open-sourced multi-speaker TTS models, the commercial APIs produce higher quality speech which is crucial for the following experiments. 8 different speaker voices are used for training set synthesis, 1 voice for validation, and 1 voice for test set. There is no voice overlap between the 3 parts. In the rest of this chapter, the synthetic data will be referred to as OOV training, OOV dev, and OOV test set respectively.

### 8.3.4 Data Augmentation

To avoid overfitting, we perturb speech on speed with the factors of 0.9, 1.0, and 1.1. Besides, the clean speech is augmented with 5 kinds of room impulse response[9]. Then, 10 noise sources [158], such as announcements, appliances, and traffic, are added to the reverberation speech on 5 levels of speech to noise ratio (0, 4, 8, 12, 16 and 20), which results in totally 10 times increase compared to original clean speech. Moreover, SpecAugment [141] with 2 frequency masks (maximum width 50) is utilized on the fly during training.

### 8.3.5 Evaluation Metrics

We use 3 metrics to evaluate experimental results of our proposed method:

- **WER**: word error rate is the ratio of error terms, i.e., substitutions, deletions, and insertions, to the total number of words in reference.
- **Recall**: recall is the number of true positives $TP$ over the sum of the number of true positives and the number of false negatives $FN$.

$$Recall = \frac{TP}{TP + FN} \tag{8.14}$$

[4]https://ai.baidu.com/tech/speech/tts
[5]https://cloud.google.com/text-to-speech
[6]https://www.xfyun.cn/doc/tts/online_tts/API.html
[7]https://cloud.tencent.com/document/api
[8]https://www.alibabacloud.com/help/doc-detail/84435.htm?spm=a2c63.p38356.b99.57.74346090w3bLLX
[9]http://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/aachen-impulse-response-database/

- **Precision**: precision is the number of true positives $TP$ over the sum of the number of true positives and the number of false positives $FP$.

$$Precision = \frac{TP}{TP + FP} \tag{8.15}$$

### 8.3.6 Training Settings

The baseline model is trained on the 960h LibriSpeech dataset with a batch size of 12, an initial learning rate of 4e-3, and warm-up steps of 25000. When doing fine-tuning for OOV experiments, we use a batch size of 15, initial learning rate of 4e-6, and anneal the learning rate with a value of 1.1 after every 3000 steps. More details about training and fine-tuning configuration can be found in Appendix D. The validation set is the mixture of the LibriSpeech dev and OOV TTS dev set. The model checkpoint performing best on the mixture validation set is used for evaluation on test sets. It is noteworthy that the attention mechanism and attention decoder are always frozen and only used for rescoring in our experiments.

## 8.4 Experimental Results

In this section, we report the experimental results from the following perspectives.

### 8.4.1 Results of Real and Synthetic Speech Mixture

To mitigate catastrophic forgetting, we mix original real speech with synthetic audio. It is still an open question what the best mixing ratio is. We fine-tune the baseline model with different ratios of real and synthetic speech mixture and report results on the standard LibriSpeech test sets (test-clean and test-other) and the synthetic OOV test set.

As shown in Table 8.2, when fine-tuning only with synthetic OOV data, the model shows the inability to retain old knowledge and performs badly on previous LibriSpeech tasks. The forgetting tendency slows down as original data incorporating into training. When the ratio of real and synthetic speech is 2:1, the model achieves the highest recall of 35.00%. The more real data is used for training, the more old task-related information is retained and the better performance is obtained on the LibriSpeech. However, the model immoderately focuses on the old tasks as the R:S ratio increases, which leads to the decrease of recall on the OOV test. We prioritize model performance regarding recall since the goal of this chapter is to enable the ASR model to learn new vocabulary and the catastrophic

forgetting issue will be tackled in the next section. Therefore, the 2:1 mixture ratio is used in the following experiments.

Table 8.2: The influence of the ratio of real and synthetic speech on ASR and OOV word learning. R and S are short for real and synthetic respectively.

| Model | R:S | WER (%) | | Recall (%) | Precision (%) | |
|---|---|---|---|---|---|---|
| | | test-clean | test-other | OOV test | | |
| **Baseline** | - | 3.18 | 8.72 | 12.53 | 1.20 | 100 |
| **Fine-tuning** | 0:1 | 39.49 | 58.42 | 45.19 | 16.20 | 94.59 |
| | 1:1 | 27.25 | 41.45 | 33.12 | 28.40 | 96.30 |
| | 2:1 | 17.26 | 30.03 | 23.27 | 35.00 | 98.13 |
| | 3:1 | 10.85 | 21.13 | 18.56 | 34.00 | 98.84 |
| | 4:1 | 6.33 | 15.34 | 16.31 | 33.80 | 99.56 |

### 8.4.2 Results of Loss Rescaling at Sentence Level

In this section, we explore the effect of loss rescaling at the sentence level. As shown in Table 8.3, using L2 and EWC regularization can efficiently overcome models suffering catastrophic forgetting and improve the recall rate on OOV with only rarely WER decrease on test-clean and test-other. We find the best $\lambda$ weight to balance the L2/EWC loss and the ASR loss is 5e7. A similar approach has been investigated by Zheng et al. [222].

All words are treated fairly when just fine-tuning the base model with synthetic audio, which leads the model to hardly focusing on the OOV words we concern. Therefore, we propose loss rescaling and encourage the model to pay more attention on OOV words by enlarging the loss of sentences containing unknown words. 100, 1000, and 10000 times loss rescaling are explored. As we can see, in Table 8.3, the OOV recall rapidly increases when rescaling the target sentences by 100 times (51.20%) compared to only fine-tuning (27.40%) using L2.

As the bigger loss weight is used, the recall further rises, but the WER on test-clean and test-other is getting a lot worse. We analysize that directly rescaling the entire sentence loss may also enhance irrelevant words or noises, which leads to gradient explosion during training and accelerates forgetting previous knowledge. Hence, we have to use a very small learning rate and clip the gradients over 2.0 to ensure the progress of fine-tuning. In contrast to L2 regularization, EWC can provide more stable and resilient protection for the weights important for the old

tasks but still get relatively high loss for the ASR performance, 20.8% on test-clean and 14.2% on test-other.

Table 8.3: Loss rescaling at sentence level with L2/EWC regularization. The values in brackets are the relative increase (↑) or decrease (↓). $\mu$ and $\lambda$ are the hyper-parameters in Eq. (8.1) and Eq. (8.12)/Eq. (8.13) respectively.

| | **Loss weight $\mu$** | **$\lambda$ weight** | **WER (%)** | | | **Recall (%)** | **Precision (%)** |
|---|---|---|---|---|---|---|---|
| **Dataset** | - | - | **test-clean** | **test-other** | **OOV test** | | |
| **Baseline** | 1 | 0 | 3.18 | 8.72 | 12.53 | 1.20 | 100 |
| **L2** | 1 | 5e7 | 3.20(↑0.6) | 8.78(↑0.7) | 12.03(↓3.9) | 27.40 | 98.56 |
| | 100 | 5e7 | 3.27(↑2.8) | 8.93(↑2.4) | 10.36(↓17.3) | 51.20 | 95.22 |
| | 1000 | 5e7 | 3.55(↑11.6) | 9.31(↑6.8) | 9.23(↓26.3) | 72.40 | 91.84 |
| | 10000 | 5e7 | 4.02(↑26.4) | 10.23(↑17.3) | 9.18(↓26.7) | 75.80 | 85.75 |
| **EWC** | 1 | 5e7 | 3.18(↓0.0) | 8.74(↑0.2) | 11.91(↓4.9) | 32.20 | 98.13 |
| | 100 | 5e7 | 3.15(↓0.9) | 8.84(↑1.4) | 10.11(↓19.3) | 60.80 | 95.02 |
| | 1000 | 5e7 | 3.46(↑8.8) | 9.07(↑4.0) | 9.20(↓25.6) | 73.60 | 91.13 |
| | 10000 | 5e7 | 3.84(↑20.8) | 9.96(↑14.2) | 9.12(↓27.2) | 76.60 | 83.75 |

### 8.4.3 Results of Loss Rescaling at Word Level

Instead of enhancing the entire sentence loss, in this section, we report the results of only rescaling unknown words. As shown in Table 8.4, the $\lambda$ weight needed to balance the ASR and L2/EWC loss scale is relatively smaller (1e7) than the one used at the sentence level (5e7) in Table 8.3 and we do not observe gradient explosion during training unless the loss weight is super large, e.g. 1e4. Using 10 times smaller loss weight can obtain the similar or even higher recall at word level rescaling, for example, using a loss weight of 100 gets a 76.4% recall rate at word level compared to using a loss weight of 1000 gets a 73.6% recall rate at sentence level when regularizing models with EWC. Additionally, rescaling loss on OOV words obtains lower WER on the standard LibriSpeech test sets. Moreover, the bigger loss weight is used, the higher recall is achieved for the OOV test. Besides, the worse WER is obtained for the LibriSpeech benchmark, which is observed at sentence level loss rescaling as well. To make a tradeoff between WER and Recall, 100 times loss rescaling is desirable with only a 2.2%/2.6% relative WER increase on test-clean/test-other and a 76.40% recall rate on OOV test.

Table 8.4: Loss rescaling at word level with L2/EWC regularization. The values in brackets are the relative increase (↑) or decrease (↓). $\mu$ and $\lambda$ are the hyper-parameters in Eq. (8.8) and Eq. (8.12)/Eq. (8.13) respectively.

| | **Loss weight $\mu$** | **$\lambda$ weight** | **WER (%)** | | | **Recall (%)** | **Precision (%)** |
|---|---|---|---|---|---|---|---|
| **Dataset** | - | - | **test-clean** | **test-other** | | **OOV test** | |
| **Baseline** | 1 | 0 | 3.18 | 8.72 | 12.53 | 1.20 | 100 |
| **L2** | 1 | 1e7 | 3.20(↑0.6) | 8.75(↑0.3) | 12.05(↓3.8) | 27.20 | 99.35 |
| | 10 | 1e7 | 3.27(↑2.8) | 8.83(↑1.3) | 10.54(↓15.9) | 52.60 | 92.75 |
| | 100 | 1e7 | 3.41(↑7.2) | 9.01(↑3.3) | 9.51(↓24.1) | 73.80 | 88.27 |
| | 1000 | 1e7 | 3.98(↑25.2) | 10.70(↑22.7) | 9.72(↓22.4) | 74.20 | 74.24 |
| **EWC** | 1 | 1e7 | 3.18(↑0.0) | 8.78(↑0.5) | 11.95(↓4.6) | 33.00 | 98.46 |
| | 10 | 1e7 | 3.20(↑0.6) | 8.83(↑1.3) | 10.34(↓17.5) | 64.40 | 91.18 |
| | 100 | 1e7 | 3.25(↑2.2) | 8.95(↑2.6) | 9.25(↓26.2) | 76.40 | 88.13 |
| | 1000 | 1e7 | 3.62(↑13.8) | 9.63(↑10.4) | 9.69(↓22.7) | 79.20 | 72.09 |

## 8.5 Summary

In this chapter, we present the use of synthetic speech to boost the ASR model on the recognition of OOV words, for instance, trending words or new named entities. Instead of just fine-tuning with audio containing OOV words, we propose to rescale loss at sentence level or word level, which encourages models to pay more attention to unknown words. Experimental results reveal that fine-tuning the baseline ASR model combined with loss rescaling and L2/EWC regularization can significantly improve OOV word recall rate and efficiently overcome models suffering catastrophic forgetting. Furthermore, loss rescaling at the word level is more stable than at the sentence level and results in less ASR performance loss on general non-OOV words and old tasks.

# Chapter 9

# Conclusion and Future Work

In this chapter, we give the answers to the research questions asked at the beginning of this thesis, followed by some potential directions for future work.

## 9.1 Answers to Research Questions

**Q1: How can neural networks reconstruct intelligible speech from mouth or face movement sequences when speech signals are noisy or not available?**

To answer this question, we conduct speech reconstruction experiments in Chapter 2 and 3. A novel encoder-decoder with attention architecture, LipSound, is proposed to map the sequence of mouth movement images directly to Mel-spectrogram to reconstruct the speech-relevant information. The speaker-dependent evaluation results demonstrate that our proposed model can generate quality Mel-spectrograms and intelligible speech, some generated audio can be found on the web[1].

In Chapter 3, we find that the large-scale crossmodal self-supervised pre-training can significantly benefit speech reconstruction in both generalizability (speaker-independent) and transferability (Non-Chinese to Chinese). In comparison to previous work, we achieve state-of-the-art results on speech quality and intelligibility in English speaker-dependent settings. Moreover, we also verify the possibility of inferring unseen speakers' voices by self-supervised pre-training with our LipSound2 model. The demo video can be found on the website[2].

[1] https://soundcloud.com/user-612210805/sets/video-to-mel
[2] https://leyuanqu.github.io/LipSound2/

**Q2: How can we improve the speech separation/recognition performance with a face image?**

We design experiments on target speech separation, where a target speaker voice is required as an auxiliary input to isolate the expected voice from a speech mixture. We replace voice references with corresponding face embeddings that are extracted from FaceNet for the face recognition task. The experimental results show that using face embeddings as additional inputs can achieve a competitive source to distortion ratio with the model using voice references on speech separation experiments. In addition, when feeding the separated speech guided by face embeddings into an ASR model, we obtain a similar word error rate with the result using voice references, which reveals that the neural network can successfully learn voice-related identity information from a face image. Besides, the quality of separated speech can be further improved by combining both face and voice references. When checking the generated audio, we find that face information can help voice embeddings to distinguish two similar voices but with different skin colors.

**Q3: How can we incorporate domain knowledge into end-to-end architectures to improve the robustness of ASR systems?**

In Chapter 5, we present two approaches to combine domain knowledge, i.e. AFs, into end-to-end learning. First, fine-tuning neural networks are proposed to concatenate hidden layer outputs of AF extractors as inputs to another RNN for ASR. Second, a progressive neural network with lateral connections from AF extractors is proposed to integrate articulatory knowledge into an end-to-end architecture. Results show that both approaches can effectively incorporate articulatory information into end-to-end learning. Furthermore, the progressive neural network brings a significant improvement compared to the baseline system and previous works.

**Q4: How can we keep ASR models robust when adding OOV words?**

In Chapter 6, we show using synthetic speech to boost the ASR model on the recognition of OOV words, for instance, trending words or new named entities. Instead of just fine-tuning with audio containing OOV words, we propose to rescale losses at sentence level or word level, which encourages models to pay more attention to unknown words. The sentence level loss rescaling is conducted on a path loss, while the word level method locates the OOV words in candidate hypotheses and enhances the word gradients during back-propagation. Experimental results reveal that fine-tuning the baseline ASR model combined with loss rescaling and L2/EWC regularization can significantly improve the OOV word recall rate and efficiently overcome models suffering catastrophic forgetting. Furthermore, loss

rescaling at the word level is more stable than at the sentence level and results in less ASR performance loss on general non-OOV words and old tasks.

## 9.2 Future Work

Although we have made substantial progress on speech reconstruction in controlled environments, there is still a significant gap in the requirements for real-world scenarios. Future work will focus on more realistic configurations, such as the variety of light conditions, moving head poses and different background environments. Moreover, the current lip reading experiments are separately conducted in two steps in which the error generated in the first step (video-to-wav) will be propagated to the second step (wav-to-text). How to jointly train the two tasks in an end-to-end fashion could be another future research direction. Besides, we are also interested in integrating our LipSound2 model into active speaker detection, speech enhancement and speech separation tasks to boost the performance of speech recognition systems in human-robot interaction.

The face embedding used in our thesis is extracted from a model mainly trained on frontal faces (VGGFace2) which is sensitive to the profile views of faces, as indicated in Figure 6.4. Future work will focus on adding faces from different angles to the face embedding net training. It is also possible to learn the face embeddings via crossmodal distillation [12] in which the voice embedding net transfers its knowledge to the face embedding net. This can be applied to scenarios where no voice embedding is available, for instance, a lecture or a colloquium where the clean speaker voice reference is usually not available, but the speaker face image is accessible on a poster or website.

Different speech attributes play different roles during speech production. Future work will investigate the weighted combination approach to automatically learn the contributions of different speech attributes. Furthermore, we are interested to integrate more domain knowledge into end-to-end learning under noisy and reverberation scenarios. The integration of AF improves ASR performance while increasing computation and time complexity. Future work will also focus on jointly training different AF extractors with one network to decrease computation and time complexity.

The proposed target word loss rescaling method is simple and effective, but there is still much left to be improved. Currently, all results are evaluated on synthetic audio data which is much different from the spontaneous speech recorded in the real world. Future work could focus on real-scenario speech collecting and

model evaluation. Additionally, the current OOV word set needs to be defined in advance, and how to automatically detect and optimize OOV words is another point to be done. The trade-off between WER on universal test sets (e.g. LibriSpeech test-clean and test-other) and recall rate on the OOV set is another issue. Dynamic L2/EWC weight can be adopted to replace the fixed $\lambda$ weight used in this work since the ASR loss will decrease in the training procedure. Later in the fine-tuning, using a fixed regularization weight probably dominates model updating. Moreover, we are interested in investigating the effectiveness of our proposed method on RNN-T and attention-based encoder-decoder ASR systems. It is also worthwhile to explore our loss rescaling on some general long tail problems, for example, image classification and voice verification.

## 9.3 Conclusion

In conclusion, to improve the robustness of end-to-end ASR systems in realistic or adverse environments, in this thesis:

1. We propose LipSound/LipSound2 to directly reconstruct speech from mouth or face movement sequences when audio signals are noisy or not available. The models are trained in a self-supervised way and do not require any human annotations. The generated representations or audio can then be utilized to improve ASR performance. We achieve significant improvement in the generated speech quality and speech intelligibility. Besides, we build a new benchmark on Chinese speech reconstruction in speaker-dependent and speaker-independent cases.

2. We verify the possibility of replacing the voice references with face embeddings in target speech separation systems. The conventional system requires a pre-enrolled voice to guide models in outputting expected clean speaker voices while our face-driven model can capture a face as auxiliary information on the fly, which enables the proposed system to easily generalize to new devices. The experimental results suggest that the face-based speech separation model can achieve competitive performance with the model using voices on both speech separation and speech recognition tasks. Moreover, the face and voice references are complementary and combining the two modalities can accomplish better performance.

3. To integrate domain knowledge into end-to-end training to increase the model robustness, a novel progressive neural network and the fine-tuning technique

are used to connect articulatory features with a CTC-based ASR system. Results show that both approaches can effectively help ASR modeling. The progressive neural network brings a significant improvement compared to the baseline system and previous works.

4. Lastly, to help ASR models learn new named entities and trending words, we propose word level and sentence level loss rescaling methods which encourage models to pay more attention to unknown vocabulary. Experimental results reveal that fine-tuning the baseline ASR model combined with loss rescaling and EWC regularization can not only significantly improve the OOV word recall rate and but also efficiently overcome models suffering catastrophic forgetting. Furthermore, loss rescaling at the word level is more stable than at the sentence level and results in less ASR performance loss on general non-OOV words and old tasks.

In summary, this thesis contributes to the field of robust speech recognition. We hope our LipSound and LipSound2 systems can inspire more researchers from computer vision and signal processing to work on multi-modal or cross-modal self-supervised learning. In addition, we also focus on practical applications, for example, our proposed face-guided speech separation system is easy to deploy on devices, for instance, smart speakers or robots. Besides, the loss rescaling technique can boost the development of end-to-end ASR systems in commercial applications since the OOV words issue is one of the most challenging problems in real-world scenarios.

# Appendix A

# Phoneme to AFs Mapping

Table A.1 shows the details of eight AF extractors (Manner, Place, Anterior, Back, Continuant, Round, Tense, Voiced). Output units mean the number of units in each AF extractor output layer. The phoneme-level transcriptions shown in the last column can be transformed into AF-level labels according to the flow diagram shown in Fig. 7.1 when building AF extractors.

Table A.1: The mapping of articulatory features and phonemes used in this thesis [212]

| AF extractor number | Output units | Category | Attribute | Phonemes |
|---|---|---|---|---|
| 1 | 39 | Manner | Vowel | iy ih eh ey ae aa aw ay ah ao oy ow uh uw er |
| | | | Fricative | jh ch s sh z zh f th v dh hh |
| | | | Nasal | m n ng |
| | | | Stop | b d g p t k |
| | | | Approximant | w y l r |
| 2 | 41 | Place | Coronal | d l n s t z |
| | | | High | ch ih iy jh sh uh uw y ow g k ng |
| | | | Dental | dh th |
| | | | Glottal | hh |
| | | | Labial | b f m p v w |
| | | | Low | aa ae aw ay oy |
| | | | Mid | ah eh ey ow |
| | | | Retroflex | er r |
| | | | Velar | g k ng |
| 3 | 14 | Others | Anterior | b d dh f l m n p s t th v z w |
| 4 | 11 | | Back | ay aa ah ao aw ow oy uh uw g k |
| 5 | 26 | | Continuant | aa ae ah ao aw ay dh eh er r ey l f ih iy oy ow s sh th uh uw v w y z |
| 6 | 10 | | Round | aw ow uw ao uh v y oy r w |
| 7 | 19 | | Tense | aa ae ao aw ay ey iy ow oy uw ch s sh f th p t k hh |
| 8 | 29 | | Voiced | aa ae ah aw ay ao b d dh eh er ey g ih iy jh l m n ng ow oy r uh uw v w y z |

# Appendix B

# Glossary of Acronyms and Abbreviations

**AFs** Articulatory Features
**ARSG** attention-based recurrent sequence generator
**ASAT** automatic speech attribute transcription
**ASR** Automatic Speech Recognition
**BRDNN** Bi-directional Recurrent Deep Neural Network
**CER** Character Error Rate
**CMLR** Chinese Mandarin Lip Reading
**CNN** Convolutional Neural Networks
**CTC** connectionist temporal classification
**DNN** Deep Neural Network
**DFSMN** Deep Feedforward Sequential Memory Network
**ESTOI** Extended Short-Time Objective Intelligibility
**FC** Fully Connected
**FST** Finite State Transducer
**GAN** Generative Adversarial Networks
**GMM** Gaussian Mixture Model
**GRU** Gated Recurrent Units
**HMM** Hidden Markov Model
**LAS** Listen, Attend and Spell
**LM** Language Model
**LPC** Linear Predictive Coding
**LRS** Lip Reading Sentence
**LRW** Lip Reading Word
**LSTM** Long-Short Term Memory

**LVCSR** Large Vocabulary Continuous Speech Recognition
**MSE** Mean Square Error
**MTCNN** Multi-task Convolutional Neural Network
**PESQ** Perceptual Evaluation of Speech Quality
**ReLU** Rectified Linear Unit
**RNN** Recurrent Neural Network
**RNN-T** Recurrent Neural Network Transducer
**SDR** Source to Distortion Ratio
**sMBR** State-level Minimum Bays Risk
**STFT** Short Time Fourier Transform
**TTS** Text-to-Speech
**WAS** Watch, Attend and Spell
**WER** Word Error Rate
**WSJ** Wall Steet Journal

# Appendix C

# Publications

- Björn Plüster, Cornelius Weber, **Leyuan Qu**, Stefan Wermter, "Hearing Faces: Target Speaker Text-to-Speech Synthesis from a Face", IEEE Automatic Speech Recognition and Understanding Workshop, 2021.
- **Leyuan Qu**, Cornelius Weber, and Stefan Wermter. LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading. arXiv preprint arXiv:2112.04748, 2021.
- **Leyuan Qu**, Cornelius Weber, and Stefan Wermter. Multimodal target speech separation with voice and face references. In Proc. INTERSPEECH, pages 1416-1420, 2020.
- Hussam Almotlak, Cornelius Weber, **Leyuan Qu**, Stefan Wermter, "Variational Autoencoder with Global and Medium Timescale Auxiliaries for Emotion Recognition from Speech", In International Conference on Artificial Neural Networks, pages 529-540. Springer, 2020.
- **Leyuan Qu**, Cornelius Weber, and Stefan Wermter. LipSound: Neural mel-spectrogram reconstruction for lip reading. In Proc. INTERSPEECH, pages 2768-2772, 2019.
- **Leyuan Qu**, Cornelius Weber, Egor Lakomkin, Johannes Twiefel, and Stefan Wermter. Combining articulatory features with end-to-end learning in speech recognition. In International Conference on Artificial Neural Networks, pages 500-510. Springer, 2018.

# Appendix D

# ASR Training Configuration

All parameters of hybrid CTC/attention ASR model used in the thesis refer to the findings and experimental results by Miao et al. [124, 123] and Zhang et al. [213].

## D.1 Configuration for Baseline Model

```
encoder: conformer
encoder_conf:
        output_size: 256
        attention_heads: 4
        linear_units: 2048
        num_blocks: 12
        dropout_rate: 0.1
        positional_dropout_rate: 0.1
        attention_dropout_rate: 0.0
        input_layer: conv2d
        normalize_before: true
        cnn_module_kernel: 31
        use_cnn_module: True
        activation_type: 'swish'
        pos_enc_layer_type: 'rel_pos'
        selfattention_type: 'rel_selfattn'

# decoder related
decoder: transformer
decoder_conf:
        attention_heads: 4
        linear_units: 2048
        num_blocks: 6
        dropout_rate: 0.1
        positional_dropout_rate: 0.1
        self_attention_dropout_rate: 0.0
        src_attention_dropout_rate: 0.0

# hybrid CTC/attention
model_conf:
        ctc_weight: 0.3
        lsm_weight: 0.1
        length_normalized_loss: false

# use raw_wav or kaldi feature
raw_wav: true

# feature extraction
collate_conf:
        # waveform level config
```

```
    wav_distortion_conf:
        wav_dither: 0
        wav_distortion_rate: 0.0
        distortion_methods: []
    speed_perturb: false
    feature_extraction_conf:
        feature_type: 'fbank'
        mel_bins: 80
        frame_shift: 10
        frame_length: 25
        using_pitch: false
    # spec level config
    # spec_swap: false
    feature_dither: 0.0
    spec_aug: true
    spec_aug_conf:
        warp_for_time: False
        num_t_mask: 2
        num_f_mask: 2
        max_t: 50
        max_f: 10
        max_w: 80

# dataset related
dataset_conf:
    max_length: 40960
    min_length: 0
    batch_type: 'static'
    batch_size: 12
    sort: true

grad_clip: 5
accum_grad: 1
max_epoch: 70
log_interval: 100

optim: adam
optim_conf:
    lr: 0.004
scheduler: warmuplr
scheduler_conf:
    warmup_steps: 25000
```

## D.2 Configuration for TTS Fine-Tuning

```
encoder: conformer
encoder_conf:
    output_size: 256
    attention_heads: 4
    linear_units: 2048
    num_blocks: 12
    dropout_rate: 0.1
    positional_dropout_rate: 0.1
    attention_dropout_rate: 0.0
    input_layer: conv2d
    normalize_before: true
    cnn_module_kernel: 31
    use_cnn_module: True
    activation_type: 'swish'
    pos_enc_layer_type: 'rel_pos'
    selfattention_layer_type: 'rel_selfattn'

# decoder related
decoder: transformer
decoder_conf:
    attention_heads: 4
    linear_units: 2048
```

```
        num_blocks: 6
        dropout_rate: 0.1
        positional_dropout_rate: 0.1
        self_attention_dropout_rate: 0.0
        src_attention_dropout_rate: 0.0

# hybrid CTC/attention
model_conf:
        ctc_weight: 0.3
        lsm_weight: 0.1
        length_normalized_loss: false

# use raw_wav or kaldi feature
raw_wav: true

# feature extraction
collate_conf:
        # waveform level config
        wav_distortion_conf:
                wav_dither: 0
                wav_distortion_rate: 0.0
                distortion_methods: []
        speed_perturb: false
        feature_extraction_conf:
                feature_type: 'fbank'
                mel_bins: 80
                frame_shift: 10
                frame_length: 25
                using_pitch: false
        # spec level config
        # spec_swap: false
        feature_dither: 0.0
        spec_aug: true
        spec_aug_conf:
                warp_for_time: False
                num_t_mask: 0
                num_f_mask: 2
                max_t: 0
                max_f: 10
                max_w: 80


# dataset related
dataset_conf:
        max_length: 40960
        min_length: 0
        batch_type: 'static'
        batch_size: 12
        sort: true

grad_clip: 2
accum_grad: 1
max_epoch: 10
log_interval: 100

optim: adam
optim_conf:
        lr: 0.000004
scheduler: warmuplr
scheduler_conf:
        warmup_steps: 0
```

# Appendix E

# Utterance Examples of Trending Words and New Named Entities Crawled from the Internet

Note: all the example sentences are crawled from Internet but not my original words. The many different examples originate from many different webpages.

1. **Trending Words**

   - Coronasomnia noun [U]
     - Meaning: the condition of being unable to sleep because of anxiety related to the coronavirus pandemic.
     - Examples:
       * Coronasomnia is the term used for sleep problems related to the pandemic.
       * This disruption is due to increased stress and anxiety, leading to what some sleep experts are calling "coronasomnia."
       * Check out this gallery to learn all about coronasomnia and how to handle it.
       * The effects of coronasomnia can range from milder disruptions, like more restlessness or feeling unrefreshed even after sleep, to more serious ones, like being unable to sleep at all.
       * Most of the information sleep experts have on coronasomnia is anecdotal, but there is plenty of it.

    - As if all the COVID fatigue and anxiety were not enough, there's another reason for coronasomnia: Our normal routines have been ripped apart.
    - Coronasomnia is a series of vicious circles.
    - The conversation around mental health during the pandemic has already begun to open up, but the related topic of insomnia brought on by COVID-19, or coronasomnia is receiving less attention.
    - The stress, disrupted routines, and unpredictability of a global pandemic is to blame for Coronasomnia, an uptick in anxiety and depression.
    - Are there certain signs that physicians should watch for that would indicate a patient may be suffering from coronasomnia?

- Maskne noun [U]
  - Meaning: acne caused or made worse by wearing a mask.
  - Examples:
    - If you notice signs of maskne, you should work with your dermatologist to create a care plan as soon as possible.
    - Maskne is a form of acne mechanica that causes breakouts in the areas covered by a face mask – the jaw, cheeks, nose, chin, and around the mouth.
    - Numerous factors may lead to the development of maskne.
    - Your first step should be to contact your dermatologist to schedule a checkup and discuss your treatment and prevention options, and generally, to ensure maskne is your concern and not something else.
    - In addition to working with your dermatologist to develop a care plan, you should keep the following maskne prevention tips in mind.
    - Skincare treatments with retinols and acids can be a great part of a treatment plan for acne-prone skin, but for people with maskne, the effects of these products are likely to be amplified under the face mask.
    - What Strategies Should I Use to Combat Maskne at Home?
    - When is it Time to See a Dermatologist for Maskne Treatment?

    - Although face masks help keep individuals safe from the novel coronavirus, they can cause a condition known as maskne.
    - This article explores maskne and its potential causes. It also suggests how individuals may treat and prevent possible skin irritation.
    - The term "maskne" was originally a reference to the development of acne after wearing a face covering or mask.

- Brexit noun [U]
  - Meaning: an exit (= act of leaving) by the United Kingdom from the European Union (short for "British exit").
  - Examples:
    - Is this finally the end of having to hear about Brexit?
    - Free trade agreements like the Brexit deal often include level playing field measures.
    - Brexit was the nickname for "British exit" from the EU, the economic and policy union that the U.K. had been a member of since 1973.
    - The UK has been bogged down by Brexit for over three years and the public is increasingly sick of hearing about it.
    - Britain will begin an 11-month transition in which it continues to abide by the bloc's rules and regulations while deciding what sort of Brexit to pursue.
    - Brexit advocates had saved for another day the tangled question of what should come next.
    - Most voters in England and Wales supported Brexit, particularly in rural areas and smaller cities.
    - With some regularity, major businesses have announced that they are leaving Britain because of Brexit, or have at least threatened to do so.
    - But opposition lawmakers and rebels in his own party seized control of the Brexit process, and moved to block a no-deal withdrawal, which would have meant Britain leaving without being able to cushion the blow of a sudden divorce.
    - Britain's exports to Europe collapsed in January as companies grappled with new terms of trade following Brexit.

    - The countdown to Brexit continues as the October 31st 2019 deadline fast approaches.

- Rollable adj.
  - Meaning: used to describe a mobile phone whose screen can be expanded into the size of a tablet.
  - Examples:
    - The world's first rollable TV, a bottom freezer refrigerator, air conditioner and dryer.
    - Samsung recently obtained a rollable display phone patent, even though the Galaxy Fold foldable smartphone isn't doing so well.
    - Sony was reported to be working on a rollable smartphone that might be released later this year or earlier next year.
    - The company also envisions smartphones with rollable and stretchable displays in the future.
    - LG Rollable would feature a flexible OLED display from Chinese firm BOE.
    - Most likely it won't and that means we will never see the LG Rollable in real again.
    - With LG no longer in the race, it now remains to be seen who comes up with a market-ready rollable smartphone first.
    - The application shows that Samsung will include all rollable phones within the same series.
    - Right now Oppo X 2021 is the only rollable phone on the "market", even though it's just a concept, which has yet to see the light of day.
    - New evidence suggests Samsung could be working on a rollable phone.
    - The company has filed for a trademark with the European Union Intellectual Property Office (EUIPO), possibly revealing the name for its first rollable smartphone.

- Blursday noun [U]
  - Meaning: a humorous way of referring to any day of the week in the time of the covid-19 pandemic, from the fact that it is sometimes difficult to know which day it is.
  - Examples:

    - Blursday can leave us feeling unmoored and spacey.
    - Alternatively, Blursday can make you feel dizzyingly productive and overwhelmed.
    - I'll also suggest a mindful reframe of Blursday.
    - "Blursday" is a popular word to use when you don't know what day of the week it is.
    - Since then, others have pointed out how days seem to run together in a merging of minutes dubbed "Blursday".
    - Although the pandemic didn't create "Blursday", it understandably gained traction through our collective experience of quarantine.
    - From "Zoom" to "social distancing" to "Blursday," how we speak and think of our common vocabulary have changed in the year.
    - With every day Blursday, it's hard to keep track of time or make memories.
    - Blursday is the term people are using for the pandemic-induced time warp that Rader and so many others have been stuck in.
    - Up to 1,000 people across the country will be selected to receive a package of pre-cooked bacon to boost their morale and make their "Blursday" a bit better.
    - What is your favour recipe using bacon to enjoy during this blursday season?

2. **New Named Entities**

- Tiktok noun [U]
  - Examples:
    - The TikTok Story begins during a moment of fraught relations between the U.S. and China.
    - After Joe Biden moved into the White House, his administration put an official hold on the TikTok ban and sale process.
    - Meanwhile, TikTok became the most downloaded app in the world, according to market research firm App Annie.
    - It has emerged that pupils who have already sat the "exams" have been sharing information on the Tiktok app.

    - I am technically too old for TikTok, sitting just outside its core 13-24-year-old market.
    - The art of the TikTok comedy sketch is unique.
    - But what has been a distinct component of TikTok culture this year has been the virality of unknown singers and artists who've been lifted from obscurity to views and listens in their millions.
    - There's a bold internet culture on TikTok.
    - I'm sure that brands soon will try to overwhelm TikTok with similar ideas.
    - Now TikTok is turning even more people into content creators via their phones.
    - With an app like TikTok, your funny video has as much chance of going around the world as the next person's.

- Instagram noun [U]
  - Examples:
    - People have now raised over $5 billion for nonprofits and personal causes through fundraisers on Facebook and Instagram.
    - Instagram is working on update that will let users post from its website.
    - The update brings the post creator to the web version of Instagram.
    - It's another way to express yourself on Instagram and the feature is now available in a few countries today with plans for more.
    - In June, Instagram will host its first ever professional development program for creators.
    - Instagram expects around 5,000 U.S.-based creators to attend Creator Week.
    - Instagram has plans to host Creator Week programs for the French and European markets as well.
    - Instagram has announced several new product updates that it says are intended to protect its younger users.
    - Going forward, Instagram will restrict the ability for adults to send direct messages to users who are under the age of 18 that don't already follow them.
    - This is because even though Instagram requires users to be at least 13 years old to create an account, many users lie.

    - Instagram is also developing other tools aimed at protecting teen users.

- Spotify noun [U]
  - Examples:
    - Spotify announces offline music, playlist and podcast downloads on Apple Watch.
    - Now if you're a Spotify user you no longer need your phone to listen to music offline.
    - The new Spotify app on Apple Watch is rolling out to all users globally over the coming weeks.
    - Also make sure you are also running the latest version of Spotify on your iPhone.
    - You're still able to use your Apple Watch to control playback from other devices using Spotify Connect.
    - Partnering with Spotify will make amazing audiobook experiences and exciting authorships easier than ever to access for our customers.
    - we will also be tapping into the opportunity of reaching new audiences who are on Spotify today, but have not yet experienced the magic of audiobooks.
    - Spotify Opens Doors for More Underrepresented Podcasters Through New Sound Up Programs.
    - I'm excited to collaborate with Spotify on this intimate concert experience that will feature songs from across my catalog.
    - we are working hard to bring the concert experience to your home, thanks to Spotify.
    - We're excited to be a part of this new initiative with Spotify that will give fans a great way to connect with their favorite artists.

- Duolingo noun [U]
  - Examples:
    - Duolingo has grown into one of the most successful CMU spinoffs and Pittsburgh tech companies.
    - Popular language-learning platform Duolingo rolled out a Yiddish course for English speakers this week.

    - The Duolingo course will cover everything from vocabulary and grammar to key cultural phrase.
    - We hope that such people will do the Duolingo course out of curiosity.
    - Duolingo is available for free on iOS, Android and the Web, with an upgrade to Duolingo Plus for $7 a month offering lessons free of ads.
    - Duolingo offers 100 total courses across nearly 40 distinct languages, from the most spoken to lesser-spoken languages.
    - Duolingo lessons adapt to your learning style.
    - A study has shown that 34 hours of Duolingo are equal to 1 university semester of language courses.
    - Learning a language on Duolingo is completely free, but you can remove ads and support free education with Plus.
    - At Duolingo, we're building new tools to give you the confidence to know what to say and how to say it.
    - In the past, Duolingo has not saved or used learner speech, but this data contains a lot of information about how second language learners actually learn speaking skills.

- Paypal noun [U]
  - Examples:
    - To assist with on-the-ground relief efforts, PayPal is donating $1 million to relief organizations in India.
    - PayPal has remained at the forefront of the digital payment revolution for more than 20 years.
    - Guided by a mission to build an inclusive economy for all people, 2020 was a pivotal year for PayPal to take action.
    - PayPal took swift and consistent action to support its stakeholders during the COVID-19 pandemic.
    - PayPal is also committed to supporting community relief efforts through new employee giving programs, fundraising campaigns and donations.
    - PayPal was a founding member of the Time to Vote movement focused on encouraging U.S. companies to give their employees time off to vote.

* In 2020, the PayPal platform enabled more than 50 million donors to contribute nearly $17 billion to more than one million nonprofits, schools, campaigns and crowdfunders.
* In 2020, PayPal made significant progress towards its goals and matched 98% of the energy in its data centers with renewable generation.
* The PayPal platform is empowering more than 375 million consumers and merchants in more than 200 markets to join and thrive in the global economy.
* PayPal Continues to Stand for Equality and Democracy.
* All transactions are settled in USD and converted to the applicable currency for the business at the standard PayPal conversion rates.

# Appendix F

# Acknowledgement

My first and big appreciation goes to my supervisor Prof. Stefan Wermter who gives me the valuable opportunity to study at the University of Hamburg and offers me insightful suggestions and full support during the doctoral study. Also, I would like to express my heartiest gratitude to Dr. Cornelius Weber for his help not only on the academic side but also the influence on my thought.

Many thanks to all of the members at the Knowledge Technology Group, particularly Egor Lakomkin, Johannes Twiefel, Sven Magg, Xiaomao Zhou, Henrique Siqueira, Alexander Sutherland, Mohammad Ali Zamani, Doreen Jirak, Pablo Barros, Stefan Heinrich, Manfred Eppe, Burhan Hafez, Di Fu, Matthias Kerzel, Mengdi Li, Tobias Hinz, Tayfun Alpay, Chandrakant Bothe. Furthermore, very special thanks to Katja Kösters and Erik Strahl for their administrative and technical support respectively.

I acknowledge the financial support from the China Scholarship Council (CSC) and the German Research Foundation DFG under project CML (TRR 169).

Finally, I would like to thank my parents Fanqing and Zhongqiang, my sisters Xiaojing and Xiaohua, and my girlfriend Wei for their unconditional, unequivocal love and support.

# Bibliography


[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: a comparison of models and an online application. In *Proc. INTERSPEECH*, page 3514–3518, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. In *Proc. INTERSPEECH*, pages 4295–4299, 2019.

[4] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. Lip2Audspec: Speech reconstruction from silent lip movements video. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2516–2520. IEEE, 2018.

[5] Petar Aleksic, Cyril Allauzen, David Elson, Aleksandar Kracun, Diego Melendo Casado, and Pedro J Moreno. Improved recognition of contact names in voice commands. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5172–5175. IEEE, 2015.

[6] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In *International Conference on Machine Learning*, pages 173–182, 2016.

[7] Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan. Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework. In *2007 IEEE International Conference on*

*Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages IV–873. IEEE, 2007.

[8] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[9] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[10] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. LipNet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[11] Kartik Audhkhasi, Bhuvana Ramabhadran, George Saon, Michael Picheny, and David Nahamoo. Direct acoustics-to-word models for English conversational speech recognition. In *Proc. INTERSPEECH*, pages 959–963, 2017.

[12] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

[14] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[15] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes. *Computer Speech & Language*, 46:605–626, 2017.

[16] Pascal Belin, Shirley Fecteau, and Catherine Bedard. Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3):129–135, 2004.

[17] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

[18] Julien Besle, Alexandra Fort, Claude Delpuech, and Marie-Hélène Giard. Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, 20(8):2225–2234, 2004.

[19] Jayadev Billa. Dropout approaches for LSTM based speech recognition systems. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5879–5883. IEEE, 2018.

[20] Stefan Braun, Daniel Neil, Jithendar Anumula, Enea Ceolini, and Shih-Chii Liu. Multi-channel attention for end-to-end speech recognition. *Proc. INTERSPEECH*, pages 17–21, 2018.

[21] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480, 1992.

[22] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[23] Vicki Bruce and Andy Young. Understanding face recognition. *British Journal of Psychology*, 77(3):305–327, 1986.

[24] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE, 2016.

[25] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[26] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. *arXiv preprint arXiv:2104.12671*, 2021.

[27] Weicong Chen, Xu Tan, Yingce Xia, Tao Qin, Yu Wang, and Tie-Yan Liu. Duallip: A system for joint lip reading and generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page

1985–1993, New York, NY, USA, 2020. Association for Computing Machinery.

[28] Zhuo Chen, Yi Luo, and Nima Mesgarani. Deep attractor network for single-microphone speaker separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250. IEEE, 2017.

[29] Zhuo Chen, Xiong Xiao, Takuya Yoshioka, Hakan Erdogan, Jinyu Li, and Yifan Gong. Multi-channel overlapped speech recognition with location guided speech extraction network. In *2018 IEEE Spoken Language Technology Workshop*, pages 558–565. IEEE, 2018.

[30] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015.

[31] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014.

[32] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 1022–1030, 2014.

[33] Hyeong-Seok Choi, Changdae Park, and Kyogu Lee. From inference to generation: End-to-end fully self-supervised generation of human face from speech. In *International Conference on Learning Representations (ICLR)*, 2020.

[34] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *Advances in Neural Information Processing Systems*, 28:577–585, 2015.

[35] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Proc. INTERSPEECH*, pages 1086—1090. 2018.

[36] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE, 2017.

[37] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

[38] Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. FaceFilter: Audio-visual speech separation using still images. In *Proc. INTERSPEECH*, pages 3481–3485, 2020.

[39] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969. IEEE, 2019.

[40] Adam Coates, Brody Huval, Tao Wang, David Wu, Bryan Catanzaro, and Ng Andrew. Deep learning with COTS HPC systems. In *International Conference on Machine Learning*, pages 1337–1345. PMLR, 2013.

[41] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[42] Thomas Le Cornu and Ben Milner. Reconstructing intelligible audio speech from visual speech features. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 34–42, 2015.

[43] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[44] Jeffrey Dean, Greg S Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V Le, Mark Z Mao, Marc'Aurelio Ranzato, Andrew Senior, Paul Tucker, et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, pages 1232–1240, 2012.

[45] Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani. Single channel target speaker extraction and recognition with speaker beam. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5554–5558. IEEE, 2018.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding.

In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[47] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[48] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. AISHELL-2: transforming mandarin ASR research into industrial scale. *arXiv preprint arXiv:1808.10583*, 2018.

[49] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 455–462, 2017.

[50] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *ACM Special Interest Group on Computer Graphics and Interactive Techniques*, 2018.

[51] Ariel Ephrat and Shmuel Peleg. Vid2Speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017.

[52] Ruchao Fan, Pan Zhou, Wei Chen, Jia Jia, and Gang Liu. An online attention-based model for speech recognition. In *Proc. of INTERSPEECH*, pages 4390–4394, 2018.

[53] Basi Garcia, Brendan Shillingford, Hank Liao, Olivier Siohan, Otavio de Pinho Forin Braga, Takaki Makino, and Yannis Assael. Recurrent neural network transducer for audio-visual speech recognition. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[54] Bryan Gick, Ian Wilson, and Donald Derrick. *Articulatory Phonetics*. John Wiley & Sons, 2012.

[55] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[56] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[57] Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori. Face2Speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image. *Proc. INTERSPEECH*, pages 1321–1325, 2020.

[58] Alex Graves. Supervised sequence labelling with recurrent neural networks. In *Studies in Computational Intelligence*, 2008.

[59] Alex Graves. Sequence transduction with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, 2012.

[60] Alex Graves. Generating sequences with recurrent neural networks. In *arXiv preprint arXiv:1308.0850*, 2013.

[61] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376, 2006.

[62] Daniel Griffin and Jae Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[63] Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, and Dong Yu. Multi-modal multi-channel target speech separation. *arXiv preprint arXiv:2003.07032*, 2020.

[64] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. INTERSPEECH*, pages 5036–5040, 2020.

[65] Jinxi Guo, Tara N Sainath, and Ron J Weiss. A spelling correction model for end-to-end speech recognition. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5651–5655. IEEE, 2019.

[66] Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. Visual features for context-aware speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5020–5024. IEEE, 2017.

[67] Kyu J Han, Ramon Prieto, and Tao Ma. State-of-the-art speech recognition using multi-stream self-attention with dilated 1D convolutions. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 54–61. IEEE, 2019.

[68] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. ContextNet: Improving convolutional neural networks for automatic speech recognition with global context. In *Proc. INTERSPEECH*, pages 3610–3614, 2020.

[69] Awni Y Hannun, Andrew L Maas, Daniel Jurafsky, and Andrew Y Ng. First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs. *arXiv preprint arXiv:1408.2873*, 2014.

[70] Mary Harper. The automatic speech recognition in reverberant environments (ASpIRE) challenge. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 547–554. IEEE, 2015.

[71] Naomi Harte and Eoin Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015.

[72] Tomoki Hayashi, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramon Astudillo, and Kazuya Takeda. Back-translation-style data augmentation for end-to-end ASR. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 426–433. IEEE, 2018.

[73] Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth Workshop on Statistical Machine Translation*, pages 187–197, 2011.

[74] Martin Heckmann, Kristian Kroschel, Christophe Savariaux, and Frédéric Berthommier. DCT-based video features for audio-visual speech recognition. In *Seventh International Conference on Spoken Language Processing*, 2002.

[75] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[76] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[77] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[78] Takaaki Hori, Shinji Watanabe, and John R Hershey. Joint CTC/attention decoding for end-to-end speech recognition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 518–529, 2017.

[79] Takaaki Hori, Shinji Watanabe, and John R Hershey. Multi-level language modeling and decoding for open vocabulary end-to-end speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 287–293. IEEE, 2017.

[80] Kyuyeon Hwang and Wonyong Sung. Character-level incremental speech recognition with recurrent neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5335–5339. IEEE, 2016.

[81] Keith Ito and Linda Johnson. The LJ speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[82] Jesper Jensen and Cees H Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

[83] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning*, pages 2342–2350. PMLR, 2015.

[84] Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1023–1027. IEEE, 2018.

[85] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigne. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4):187–207, 1999.

[86] Shreya Khare, Rahul Aralikatte, and Senthil Mani. Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization. In *Proc. INTERSPEECH*, pages 3208–3212, 2018.

[87] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, 2018.

[88] Chanwoo Kim and Richard M Stern. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(7):1315–1329, 2016.

[89] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4835–4839. IEEE, 2017.

[90] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[91] Simon King and Paul Taylor. Detection of phonological features in continuous speech using neural networks. *Computer Speech & Language*, 14(4):333–353, 2000.

[92] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

[93] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

[94] Katrin Kirchhoff. Robust speech recognition using articulatory information. In *Ph. D. thesis*. Bielefeld University, 1999.

[95] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[96] Kate M Knill, Mark JF Gales, Shakti P Rath, Philip C Woodland, Chao Zhang, and S-X Zhang. Investigation of multilingual deep neural networks for spoken term detection. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 138–143. IEEE, 2013.

[97] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[98] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE, 2017.

[99] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018.

[100] Alexandros Koumparoulis and Gerasimos Potamianos. MobiLipNet: Resource-efficient deep learning based lipreading. In *Proc. INTERSPEECH*, pages 2763–2767, 2019.

[101] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE, 2020.

[102] Yaman Kumar, Rohit Jain, Khwaja Mohd Salik, Rajiv Ratn Shah, Yifang Yin, and Roger Zimmermann. Lipper: Synthesizing thy speech using multi-view lipreading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2588–2595, 2019.

[103] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite bert for self-supervised learn-

ing of language representations. In *International Conference on Learning Representations (ICLR)*, 2019.

[104] Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. You do not need more data: improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444. IEEE, 2020.

[105] Thomas Le Cornu and Ben Milner. Generating intelligible audio speech from visual speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9):1751–1761, 2017.

[106] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324. IEEE, 1998.

[107] Chin-Hui Lee, Mark A Clements, Sorin Dusan, Eric Fosler-Lussier, Keith Johnson, Biing-Hwang Juang, and Lawrence R Rabiner. An overview on automatic speech attribute transcription (asat). In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[108] Li Lee and Richard Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing*, 6(1):49–60, 1998.

[109] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2019.

[110] Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde. Jasper: An end-to-end convolutional neural acoustic model. In *Proc. INTERSPEECH*, pages 71—75, 2019.

[111] Jinyu Li, Guoli Ye, Rui Zhao, Jasha Droppo, and Yifan Gong. Acoustic-to-word model without OOV. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 111–117. IEEE, 2017.

[112] Tatiana Likhomanenko, Gabriel Synnaeve, and Ronan Collobert. Who needs words? lexicon-free speech recognition. In *Proc. INTERSPEECH*, pages 3915–3919, 2019.

[113] Jindong Liu, David Perez-Gonzalez, Adrian Rees, Harry Erwin, and Stefan Wermter. A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation. *Neurocomputing*, 74(1-3):129–139, 2010.

[114] Mingshuang Luo, Shuang Yang, Shiguang Shan, and Xilin Chen. Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020.

[115] Yi Luo and Nima Mesgarani. TasNet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2018.

[116] Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019.

[117] Yi Luo and Nima Mesgarani. Separating varying numbers of sources with auxiliary autoencoding loss. In *Proc. INTERSPEECH*, pages 2622–2626, 2020.

[118] John MacDonald and Harry McGurk. Visual influences on speech perception processes. *Perception & Psychophysics*, 24(3):253–257, 1978.

[119] Shinji Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech Production and Speech Modelling*, pages 131–149. Springer, 1990.

[120] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE, 2020.

[121] Sameer R Maskey, Michiel Bacchiani, Brian Roark, and Richard Sproat. Improved name recognition with meta-data dependent name networks. In *2004*

*IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–789. IEEE, 2004.

[122] Lauren W Mavica and Elan Barenholtz. Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance*, 39(2):307, 2013.

[123] Haoran Miao, Gaofeng Cheng, Changfeng Gao, Pengyuan Zhang, and Yonghong Yan. Transformer-based online CTC/attention end-to-end speech recognition architecture. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6084–6088. IEEE, 2020.

[124] Haoran Miao, Gaofeng Cheng, Pengyuan Zhang, Ta Li, and Yonghong Yan. Online Hybrid CTC/Attention architecture for end-to-end speech recognition. In *Proc. INTERSPEECH*, pages 2623–2627, 2019.

[125] Yajie Miao and Florian Metze. Open-domain audio-visual speech recognition: A deep learning approach. In *Proc. INTERSPEECH*, pages 3414–3418, 2016.

[126] Yajie Miao and Florian Metze. End-to-end architectures for speech recognition. In *New Era for Robust Speech Recognition*, pages 299–323. Springer, 2017.

[127] Yajie Miao, Florian Metze, and Shourabh Rawat. Deep maxout networks for low-resource speech recognition. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 398–403. IEEE, 2013.

[128] Daniel Michelsanti, Olga Slizovskaia, Gloria Haro, Emilia Gómez, Zheng-Hua Tan, and Jesper Jensen. Vocoder-based speech synthesis from silent videos. In *Proc. INTERSPEECH*, pages 3530–3534, 2020.

[129] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR)*, 2013.

[130] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[131] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33, 2020.

[132] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

[133] Niko Moritz, Takaaki Hori, and Jonathan Le Roux. Triggered attention for end-to-end speech recognition. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5666–5670. IEEE, 2019.

[134] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable PINs: Cross-modal embeddings for person identity. In *Proceedings of the European Conference on Computer Vision*, pages 71–88, 2018.

[135] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[136] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2Face: Learning the face behind a voice. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7539–7548, 2019.

[137] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. In *ISCA Speech Synthesis Workshop (SSW)*, 2016.

[138] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[139] Ashutosh Pandey and DeLiang Wang. A new framework for CNN-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(7):1179–1188, 2019.

[140] Dharin Parekh, Ankitesh Gupta, Shharrnam Chhatpar, Anmol Yash Kumar, and Manasi Kulkarni. Lip reading using convolutional auto encoders as feature extractor. *arXiv preprint arXiv:1805.12371*, 2018.

[141] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. SpecAugment: A simple data augmentation

method for automatic speech recognition. In *Proc. INTERSPEECH*, pages 2613–1617, 2019.

[142] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: Speech enhancement generative adversarial network. In *Proc. INTERSPEECH*, pages 3642–3646, 2017.

[143] Douglas B Paul and Janet Baker. The design for the wall street journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[144] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin. Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 36–40. IEEE, 2018.

[145] Stavros Petridis, Zuwei Li, and Maja Pantic. End-to-end visual speech recognition with LSTMs. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2592–2596. IEEE, 2017.

[146] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. In *International Conference on Learning Representations (ICLR)*, pages 214–217, 2018.

[147] Gerasimos Potamianos, Hans Peter Graf, and Eric Cosatto. An image transform approach for HMM-based automatic lipreading. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, pages 173–177. IEEE, 1998.

[148] Daniel Povey, Hossein Hadian, Pegah Ghahremani, Ke Li, and Sanjeev Khudanpur. A time-restricted self-attention layer for ASR. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878. IEEE, 2018.

[149] Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. A comparison of sequence-to-sequence models for speech recognition. In *Proc. INTERSPEECH*, pages 939–943, 2017.

[150] Rohit Prabhavalkar, Tara N Sainath, Bo Li, Kanishka Rao, and Navdeep Jaitly. An analysis of "Attention" in sequence-to-sequence models. In *Proc. INTERSPEECH*, pages 3702–3706, 2017.

[151] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805, 2020.

[152] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: A flow-based generative network for speech synthesis. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE, 2019.

[153] Leyuan Qu, Cornelius Weber, Egor Lakomkin, Johannes Twiefel, and Stefan Wermter. Combining articulatory features with end-to-end learning in speech recognition. In *International Conference on Artificial Neural Networks*, pages 500–510. Springer, 2018.

[154] Leyuan Qu, Cornelius Weber, and Stefan Wermter. LipSound: Neural mel-spectrogram reconstruction for lip reading. In *Proc. INTERSPEECH*, pages 2768–2772, 2019.

[155] Leyuan Qu, Cornelius Weber, and Stefan Wermter. Multimodal target speech separation with voice and face references. In *Proc. INTERSPEECH*, pages 1416–1420, 2020.

[156] Leyuan Qu, Cornelius Weber, and Stefan Wermter. Lipsound2: Self-supervised pre-training for lip-to-speech reconstruction and lip reading. *arXiv preprint arXiv:2112.04748*, 2021.

[157] Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-Transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 193–199. IEEE, 2017.

[158] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. A scalable noisy speech dataset and online subjective test framework. In *Proc. INTERSPEECH*, pages 1816–1820, 2019.

[159] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 2, pages 749–752. IEEE, 2001.

[160] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech recognition with augmented synthesized speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 996–1002. IEEE, 2019.

[161] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Generating synthetic audio data for attention-based speech recognition systems. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073. IEEE, 2020.

[162] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[163] Haşim Sak, Andrew Senior, Kanishka Rao, Ozan Irsoy, Alex Graves, Françoise Beaufays, and Johan Schalkwyk. Learning acoustic frame labeling for speech recognition with recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4280–4284. IEEE, 2015.

[164] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[165] Stefan R Schweinberger, David Robertson, and Jürgen M Kaufmann. Hearing facial identities. *Quarterly Journal of Experimental Psychology*, 60(10):1446–1456, 2007.

[166] Kouhei Sekiguchi, Yoshiaki Bando, Kazuyoshi Yoshii, and Tatsuya Kawahara. Bayesian multichannel speech enhancement with a deep speech prior. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1233–1239. IEEE, 2018.

[167] Imran Sheikh, Dominique Fohr, Irina Illina, and Georges Linares. Modelling semantic context of OOV words in large vocabulary continuous speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(3):598–610, 2017.

[168] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan,

et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[169] Brendan Shillingford, Yannis Assael, Matthew W Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, et al. Large-scale visual speech recognition. In *Proc. INTERSPEECH*, pages 4135–4139, 2018.

[170] Peter Livingston Silsbee and Alan C Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, 1996.

[171] Khe Chai Sim, Françoise Beaufays, Arnaud Benard, Dhruv Guliani, Andreas Kabel, Nikhil Khare, Tamar Lucassen, Petr Zadrazil, Harry Zhang, Leif Johnson, et al. Personalization of end-to-end speech recognition on mobile devices for named entities. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 23–30. IEEE, 2019.

[172] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2014.

[173] Sabato Marco Siniscalchi and Chin-Hui Lee. A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. *Speech Communication*, 51(11):1139–1153, 2009.

[174] Sabato Marco Siniscalchi, Dau-Cheng Lyu, Torbjørn Svendsen, and Chin-Hui Lee. Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):875–887, 2011.

[175] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with LSTMs for lipreading. In *Proc. INTERSPEECH*, pages 3652––3656, 2017.

[176] George Sterpu and Naomi Harte. Towards lipreading sentences with active appearance models. In *Proc. The 14th International Conference on Auditory-Visual Speech Processing*, pages 70–75, 2017.

[177] William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2):212–215, 1954.

[178] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda. Speaker-dependent wavenet vocoder. In *Proc. INTERSPEECH*, pages 1118–1122, 2017.

[179] Kwanchiva Thangthai, Helen L Bear, and Richard Harvey. Comparing phonemes and visemes with DNN-based lipreading. In *MVC Lipreading Workshop*, 2018.

[180] Kwanchiva Thangthai and Richard Harvey. Improving computer lipreading via DNN sequence discriminative training techniques. Proc. INTERSPEECH, 2017.

[181] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Local monotonic attention mechanism for end-to-end speech and language processing. In *the International Joint Conference on Natural Language Processing*, 2017.

[182] Jean-Marc Valin and Jan Skoglund. LPCNet: Improving neural speech synthesis through linear prediction. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5891–5895. IEEE, 2019.

[183] Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro. FlowTron: an autoregressive flow-based generative network for text-to-speech synthesis. In *International Conference on Learning Representations (ICLR)*, 2020.

[184] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[185] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *Proc. INTERSPEECH*, volume 2013, pages 2345–2349, 2013.

[186] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

[187] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 391–408, 2018.

[188] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Video-driven speech reconstruction using generative adversarial networks. In *Proc. INTERSPEECH*, pages 4125–4129, 2019.

[189] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE, 2018.

[190] Michael Wand, Jan Koutník, and Jürgen Schmidhuber. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6115–6119. IEEE, 2016.

[191] Michael Wand and Jürgen Schmidhuber. Improving speaker-independent lipreading with domain-adversarial training. In *Proc. INTERSPEECH*, pages 2415—2419, 2017.

[192] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[193] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno. VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Proc. INTERSPEECH*, pages 2728–2732. IEEE, 2019.

[194] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno. The conversation: Deep audio-visual speech enhancement. In *Proc. INTERSPEECH*, pages 3244–3248. IEEE, 2018.

[195] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[196] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. Complex spectral mapping for single-and multi-channel speech enhancement and robust

ASR. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1778–1787, 2020.

[197] Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, and Rita Singh. Disjoint mapping network for cross-modal matching of voices and faces. In *International Conference on Learning Representations*, 2018.

[198] Yandong Wen, Bhiksha Raj, and Rita Singh. Face reconstruction from voice using generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 5266–5275, 2019.

[199] Xinshuo Weng and Kris Kitani. Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading. *arXiv preprint arXiv:1905.02540*, 2019.

[200] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with LSTMs recurrent neural networks and its application to noise-robust ASR. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 91–99. Springer, 2015.

[201] Ian Williams, Anjuli Kannan, Petar S Aleksic, David Rybach, and Tara N Sainath. Contextual speech recognition in end-to-end neural network systems using beam search. In *Proc. INTERSPEECH*, pages 2227–2231, 2018.

[202] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989.

[203] Kevin Wilson, Michael Chinen, Jeremy Thorpe, Brian Patton, John Hershey, Rif A Saurous, Jan Skoglund, and Richard F Lyon. Exploring tradeoffs in models for low-latency speech enhancement. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 366–370. IEEE, 2018.

[204] Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, and Dong Yu. Time domain audio visual speech separation. *arXiv preprint arXiv:1904.03760*, 2019.

[205] Xiong Xiao, Zhuo Chen, Takuya Yoshioka, Hakan Erdogan, Changliang Liu, Dimitrios Dimitriadis, Jasha Droppo, and Yifan Gong. Single-channel speech extraction using speaker inventory and attention network. In *2019*

*IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 86–90. IEEE, 2019.

[206] Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. The Microsoft 2017 conversational speech recognition system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5934–5938. IEEE, 2018.

[207] Kai Xu, Dawei Li, Nick Cassimatis, and Xiaolong Wang. LCANet: End-to-end lipreading with cascaded attention-CTC. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 548–555. IEEE, 2018.

[208] Junichi Yamagishi, Heiga Zen, Tomoki Toda, and Keiichi Tokuda. Speaker-independent HMM-based speech synthesis system: HTS-2007 system for the Blizzard Challenge. 2007.

[209] Chenzhao Yang, Shilin Wang, Xingxuan Zhang, and Yun Zhu. Speaker-independent lipreading with limited data. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2181–2185. IEEE, 2020.

[210] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[211] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017.

[212] Dong Yu, Sabato Marco Siniscalchi, Li Deng, and Chin-Hui Lee. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4169–4172. IEEE, 2012.

[213] Binbin Zhang, Di Wu, Chao Yang, Xiaoyu Chen, Zhendong Peng, Xiangming Wang, Zhuoyuan Yao, Xiong Wang, Fan Yu, Lei Xie, et al. WeNet:

Production first and production ready end-to-end speech recognition toolkit. In *Proc. INTERSPEECH*, 2021.

[214] Binbin Zhang, Di Wu, Zhuoyuan Yao, Xiong Wang, Fan Yu, Chao Yang, Liyong Guo, Yaguang Hu, Lei Xie, and Xin Lei. Unified streaming and non-streaming two-pass end-to-end model for speech recognition. In *arXiv preprint arXiv:2012.05481*, 2020.

[215] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[216] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833. IEEE, 2020.

[217] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.

[218] Shiliang Zhang, Ming Lei, Zhijie Yan, and Lirong Dai. Deep-FSMN for large vocabulary continuous speech recognition. *arXiv preprint arXiv:1803.05030*, 2018.

[219] Ding Zhao, Tara N Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. Shallow-fusion end-to-end contextual biasing. In *Proc. INTERSPEECH*, pages 1418–1422, 2019.

[220] Ya Zhao, Rui Xu, and Mingli Song. A cascade sequence-to-sequence model for Chinese Mandarin lip reading. In *Proceedings of the ACM Multimedia Asia*, pages 1–6. 2019.

[221] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. Hearing Lips: Improving lip reading by distilling speech recognizers. In *AAAI*, pages 6917–6924, 2020.

[222] Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end ASR systems. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678. IEEE, 2021.

[223] Katerina Zmolikova, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani. Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. In *Proc. INTERSPEECH*, pages 2655–2659. IEEE, 2017.

[224] Geoffrey Zweig, Chengzhu Yu, Jasha Droppo, and Andreas Stolcke. Advances in all-neural speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4805–4809. IEEE, 2017.

# Erklärung der Urheberschaft

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, 12.01.2022

Ort, Datum

Unterschrift

# Erklärung zur Veröffentlichung

Ich erkläre mein Einverständnis mit der Einstellung dieser Dissertation in den Bestand der Bibliothek.

Hamburg, 12.01.2022

Ort, Datum

Unterschrift