INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Cross-Modal Knowledge Distillation Method for Automatic Cued Speech Recognition

Jianrong Wang[1], *Ziyue Tang*[2], *Xuewei Li*[1], *Mei Yu*[1], *Qiang Fang*[3], *Li Liu*[4*]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Tianjin International Engineering Institute, Tianjin University, Tianjin, China
[3]Institute of Linguistics, Chinese Academy of Social Science, Beijing, China
[4]Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong, Shenzhen, China
liuli@cuhk.edu.cn

Abstract

Cued Speech (CS) is a visual communication system for the deaf or hearing impaired people. It combines lip movements with hand cues to obtain a complete phonetic repertoire. Current deep learning based methods on automatic CS recognition suffer from a common problem, which is the data scarcity. Until now, there are only two public single speaker datasets for French (238 sentences) and British English (97 sentences). In this work, we propose a cross-modal knowledge distillation method with teacher-student structure, which transfers audio speech information to CS to overcome the limited data problem. Firstly, we pretrain a teacher model for CS recognition with a large amount of open source audio speech data, and simultaneously pretrain the feature extractors for lips and hands using CS data. Then, we distill the knowledge from teacher model to the student model with frame-level and sequence-level distillation strategies. Importantly, for frame-level, we exploit multi-task learning to weigh losses automatically, to obtain the balance coefficient. Besides, we establish a five-speaker British English CS dataset for the first time. The proposed method is evaluated on French and British English CS datasets, showing superior CS recognition performance to the state-of-the-art (SOTA) by a large margin.

Index Terms: Cued Speech, Cross-modal knowledge distillation, Teacher-student structure, Cued Speech recognition

1. Introduction

Cued Speech (CS) is a visual mode of communication for hearing impaired people proposed by Cornett [1] in 1967. It is to solve the confusion of lip reading. CS uses the shape and position of hand to assist in lip reading to make visual communication easier and more efficient for hearing impaired people. The hand shape is used to encode consonants, and the hand position is used to encode vowels (see Figure 1).

Currently, there are some automatic CS recognition works [2, 3, 4, 5, 6] based on two small scale single-speaker datasets (one is in French [7] and the other is in British English [8, 9]). The tandem architecture in [4] achieved 62% accuracy on the French CS dataset with single speaker and 238 French sentences. More recently, [5] proposed a deep sequence learning approach, which consists of an image learner based on convolutional neural networks (CNNs) [10] and a fully convolutional encoder-decoder. It was evaluated on the British English CS dataset for the first time achieving 63.75% with single speaker and 97 British English sentences. These two works were both focus on continues CS recognition with no artificial marks [2, 3, 11, 12] but their performance were still limited by the data scarcity, which leads to the overfitting in a deep neural network training.

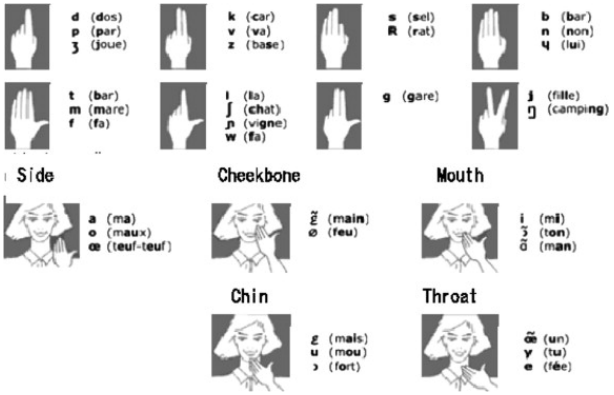

Figure 1: *Chart of the French Cued Speech (from [2]).*

Knowledge distillation with the teacher-student structure can transfer and preserve the cross-modality knowledge successfully [13, 14, 15, 16, 17, 18]. It has been successfully applied in many fields. For example, in [19], Gupta et al. transferred the knowledge from the teacher model to the student model relying on unlabeled paired samples involving both RGB and depth images. Zhao et al. [20] used synchronized radio signals and camera images to transfer the knowledge across modalities for radio-based human pose estimation. Thoker and Gall [13] transferred the knowledge obtained from RGB videos to a skeleton-based human action recognition model.

In this study, considering the limited size and multi-modality (including audios and videos) of CS data, we innovatively propose a cross-modal knowledge distillation architecture, which transfers the knowledge learned from audio speech to visual-based CS (see Figure 2). Although the size of CS dataset is limited, many acoustic speech datasets are publicly available. What's more, CS and audio speech are both phoneme-level coding thus they have the same phoneme semantics. It motivates us to exploit cross-modal distillation in CS recognition. Firstly, we pretrain a teacher model for phoneme recognition with a large amount of open source audio speech data, and simultaneously pretrain feature extractors for lips, hand shape and hand position. Then, we distill the speech knowledge from the large teacher model into the small student model to estimate the parameters of bi-directional long short-term memory (BiLSTM) [21] in the student network with frame-level

* Corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-432

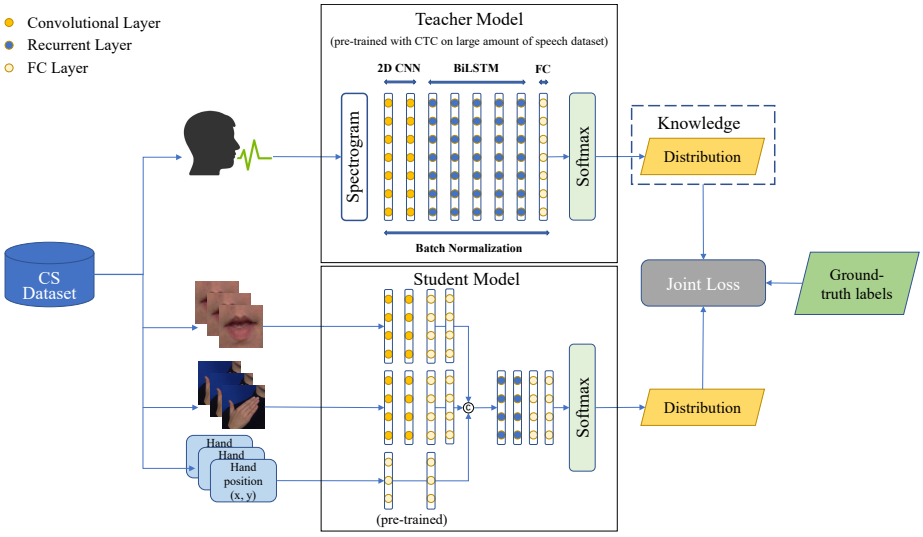

Figure 2: *Teacher-student structure for automatic CS recognition.*

and sequence-level [22, 23] distillation strategies. For frame-level distillation, the weights of losses are learned automatically based on multi-task learning, and then they are used to compute the balance coefficient (*i.e.*, hyperparameters used to balance the gradient among loss functions).

Compared with [4], we achieve significant CS phoneme recognition accuracy improvement of 12.2% in frame-level (17% in sequence-level) for French task. Compared with the state-of-the-art (SOTA) [5], we improve 3.32% in frame-level (8.12% in sequence-level) for French CS and 1.15% in frame-level (7.95% in sequence-level) for British English CS with single speaker.

In summary, the key contributions of this work are:

1. This is the first attempt to apply cross-modal knowledge distillation method to automatic CS recognition to overcome the problem of overfitting caused by limited size of CS dataset.
2. To obtain a better model automatically, we exploit multi-task learning to weigh losses, and then compute the balance coefficient to balance the gradient among loss functions.
3. We establish the first five-speaker British English CS dataset with 390 English consecutive sentences.

2. Methodology

In this section, we introduce our method, which includes teacher network and student network in the cross-modal knowledge distillation architecture, as well as frame-level and sequence-level distillation strategies.

2.1. DeepSpeech2 teacher network

DeepSpeech2 [24] is an end-to-end deep learning approach for automatic speech recognition (ASR) with good generalization performance. The architecture of DeepSpeech2 we used is shown as the Teacher Model in Figure 2. It consists of two 2D convolutional layers followed by five BiLSTM layers with batch normalization [25] and a fully connected layer (FC). It is trained with the connectionist temporal classification (CTC) [26] loss function to predict audio speech transcriptions.

To use DeepSpeech2 as teacher model for CS recognition, we translate sentence label into CS phoneme sequence. Let $x^{(i)}$ is a time-series, where each time-slice is a vector of audio features, and $y^{(i)}$ is the corresponding label. The inputs of the network are spectrogram of power normalized audio clips, and the outputs are the phonemes. At each output time-step t, BiLSTM makes a prediction over phonemes, $p(l_t|x)$ where $l_t \in \{blank, \Gamma\}$. Γ is the set of CS phonemes that are different in French and British English. We represent the CTC loss function as $L(x, y; \theta)$, where (x, y) is an input-output pair, and θ is the current parameters of the network. The network parameters are updated making use of the derivative $\nabla_\theta L(x, y; \theta)$.

2.2. CNN-BiLSTM student network

As shown in Figure 2, the architecture of the student model consists of two pre-trained CNNs and one pre-trained artificial neural network (ANN) [27], followed by two BiLSTM layers and two FC layers. CNNs act as feature extractors. Each focuses on lips or hands separately, trained with a set of 8 visemes [28] as targets for lips and 8 shapes for hands. The details of the CNNs and ANN network structure are the same as [4]. After extracting features based on CNNs and ANN, the three stream features are concatenated in a single feature vector as the input of BiLSTM.

To verify the effectiveness of knowledge distillation for CS recognition in frame-level and sequence-level, we evaluate the student model trained by minimizing Cross Entropy (CE) loss and CTC loss, respectively.

2.3. Knowledge distillation strategy

We consider two distillation strategies *i.e.*, frame-level and sequence-level. The standard knowledge distillation applied to frame-level prediction should be effective for CS recognition, and a novel sequence-level knowledge distillation strategy might further improve performance.

2.3.1. Frame-level distillation strategy

For frame-level distillation strategy, we try an existing general method firstly. Considering a generalized $softmax$ function

[29]:

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}, \quad (1)$$

where z_i is the logit from network output layer, q_i is the class probability corresponding to each z_i, and T is the temperature coefficient that is normally set to 1. Both the soft targets from the teacher and the ground-truth label are of great importance for improving the performance of the student model. Therefore, we consider Kullback-Leibler Divergence (KL) as distillation loss and CE as student loss as follow:

$$L(x_t, x_s; W) = \alpha KL(P_{(T)}, Q_{(T)}) * T^2 + (1-\alpha) CE(Y, Q_{(1)}), \quad (2)$$

where x_t is the CS audio input of the pre-trained teacher model and x_s is the input of the student model which is concatenated feature vector. W are the parameters of the student model, and α is a regulated parameter. $P_{(T)}$ and $Q_{(T)}$ are the $softmax$ output of teacher model and student model, respectively when temperature is T. Y is the one-hot ground-truth label, and $Q_{(1)}$ is the $softmax$ output of student model when the temperature is 1. For this loss function, we tune T and α manually to find a model with higher CS phoneme recognition accuracy.

The loss reduction curve is shown in the Figure 3 when T is set to 1 and α is set to 0.5. We observe that the gradient of KL is almost 0, which means that the knowledge of the teacher model is not effectively distilled to the student model, and there is an imbalance in the gradient between KL and CE. To avoid tuning T and α manually, we derive a multi-task joint loss function as Equation (3) based on the task uncertainty.

$$L_{mtl} = \frac{1}{\sigma_1^2} KL(P, Q) + \frac{1}{\sigma_2^2} CE(Y, Q) + log\sigma_1 + log\sigma_2, \quad (3)$$

where T and α are not considered anymore, and we automatically learn the observation noise scalar σ_1 and σ_2. As reported in [30], the task uncertainty captures the relative confidence between tasks, reflecting the uncertainty inherent to the regression or classification task. Therefore, in our two classification tasks (*i.e.*, the student model learns from the teacher model and learns from the ground-truth) of the distillation process, uncertainty is used as a basis to weigh losses.

Further, to balance the loss gradient between these two tasks, the loss joint function based on balance coefficient (*i.e.*, a) is proposed as:

$$L(x_t, x_s; W) = \frac{1}{2}(a \times KL(P_{(1)}, Q_{(1)}) + CE(Y, Q_{(1)})), \quad (4)$$

We obtain the balance coefficient through $a = \frac{\sigma_2^2}{\sigma_1^2}$.

2.3.2. *Sequence-level distillation strategy*

Similarly, for sequential distillation, both the soft targets from the teacher and the ground-truth label are of great importance for improving the performance of the student model. Therefore, for student network, a distillation loss that assists CTC is explored. Rather than using KL divergence, we apply cosine similarity (cos) as the distillation loss, since it pays more attention to the difference in the direction between the two vectors. It is shown as follows:

$$L(x_t, x_s; W) = \frac{1}{2}(1 - cos(S_s, S_t) + CTC(Y, logQ)), \quad (5)$$

where x_t, x_s, W and Q are the same as described in section 2.3.1. S_s and S_t are the frame-based logits before the $softmax$ layer of the student and teacher model, respectively. Y is corresponding transcript.

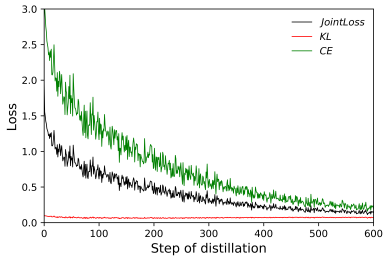

Figure 3: *Comparison of gradient descent for KL, CE and their joint loss when $T = 1$ and $\alpha = 0.5$ in Equation (2).*

3. Experiments

3.1. Datasets

CS datasets for student model We use a public French CS dataset [7] and the newly built British English CS dataset with 5 speakers specially for this work. The detail of the public single-speaker French can refer to [7]. Here, we mainly introduce the new multi-speaker British English CS dataset.

A single-speaker (named CA) British English CS dataset [8] which contains 97 sentences. Now, the dataset has been expanded with data for other four CS interpreters named EM, KA, LD and VK, respectively. There are 390 British English audios and videos of all five speakers in total. The RGB video image recorded at 25fps. In the British English CS system, the 12 monophthongs are encoded by four hand positions and the 8 diphthongs are encoded by four hand slips, while the 24 consonants are encoded by eight hand shapes.

Datasets for teacher model pre-training We train two teacher models for distillation with the French speech dataset and English speech dataset in Common Voice [31].

Common Voice is an open source project launched by Mozilla including French and English speech data. The total number of verification hours of French version is 350h and English version is 1118h. We convert the French sentence text into French CS phonemes using Lliaphon [32]. Then British English phonetic transcript is obtained by eSpeak (*i.e.*, a compact open source software speech synthesizer for English and other languages). It should be mentioned that the English datasets only contains 8% of British English, and the remaining 92% consists of American English and dialects.

3.2. Protocol and metrics

In our experiments, each CS dataset for student model uses 80% of the data for training, 10% of the data for validation, and the other 10% for testing. These three parts of data do not overlap. All models are evaluated in CS phoneme accuracy (Acc) which is defined as $Acc(\%) = 1 - PER(\%)$. PER is the phoneme error rate.

3.3. Result and analysis

We evaluate the teacher model, as well as the student model trained with distillation in frame-level and sequence-level on CS datasets.

3.3.1. *Evaluation on teacher model*

The phoneme recognition accuracy of the teacher model is 83.7% on French CS and 70.1% on the new British English CS

Table 1: *Performance comparison in Acc(%) in frame-level distillation. Fully Conv [5] is the SOTA.* JLF1 *is the best performance obtained by minimizing the joint loss function represented by Equation (2).* JLF2_{mtl} *represents the performance of the optimal model obtained by Equation (3).* JLF3 *is the performance obtained by Equation (4).*

Dataset	French	Single-speaker (British)	Multi-speaker (British)
CNN-HMM [4]	62%	—	—
Fully Conv [5]	70.88%	63.75%	—
Student CE	64.4%	52.5%	62.3%
JLF1	72.5% ($T=5, \alpha=0.5$)	61.3% ($T=5, \alpha=0.7$)	65.7% ($T=5, \alpha=0.7$)
JLF2_{mtl}	72.5% ($\sigma_1=0.37, \sigma_2=1.20$)	63.1% ($\sigma_1=0.59, \sigma_2=1.02$)	68.5% ($\sigma_1=0.23, \sigma_2=0.74$)
JLF3	**74.2%** ($a=\frac{\sigma_2^2}{\sigma_1^2}\approx 10$)	**64.9%** ($a=\frac{\sigma_2^2}{\sigma_1^2}\approx 3$)	**69.7%** ($a=\frac{\sigma_2^2}{\sigma_1^2}\approx 10$)

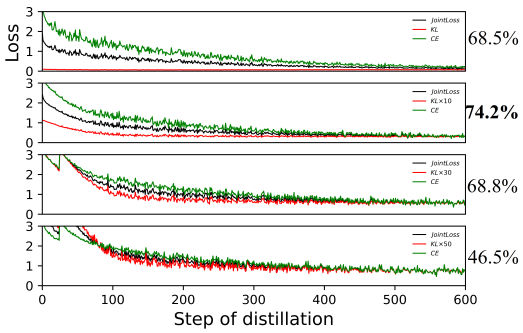

Figure 4: *Comparison of gradient descent among KL, CE and their joint loss with different balance coefficients in French CS recognition task.*

audio speech data. We hypothesize that the low performance of the English teacher model is caused by the limited British English data (only 8% in Common Voice French dataset).

3.3.2. Evaluation on student model

Frame-level distillation For frame-level knowledge distillation, we evaluate three loss functions (*i.e.*, Equation (2), (3) and (4)). In Table 1, we find out that the knowledge distillation method bring considerable improvement compared with training student model with baseline CE. On each CS dataset, JLF3 is the best result. We achieve Acc improvement of 12.2% compared with CNN-HMM [4], and 3.32% compared with SOTA [5] in French CS task. In single-speaker British CS task, we improve 1.15% Acc compared with SOTA. For the new dataset with multi speakers, JLF3 achieves Acc of 69.7%.

More balance coefficients are explored to show the effectiveness of the joint loss function we proposed (see Equation (4)). From Figure 4, it can be observed that the closer the gradient between the KL and CE at the beginning of the training step, the more efficient the distillation.

Sequence-level distillation As demonstrated in Table 2, in addition to multi-speaker, as an auxiliary loss, cos performs better than training the student model with CTC alone. The English teacher model (70.1%) is not better than the student model trained by CTC (78.6%), so that it is not enough to play a positive role in guiding the student model.

Table 2: *Performance comparison by Acc(%) in sequence-level distillation. Fully Conv [5] is the SOTA.*

Dataset	French	Single-speaker (British)	Multi-speaker (British)
CNN-HMM	62%	—	—
Fully Conv	70.88%	63.75%	—
Student CTC	77%	68.6%	78.6%
KL+CTC	75%	68.4%	75.2%
cos+CTC	**79%**	**71.4%**	77.5%

To conclude, we verify the effectiveness of knowledge distillation, the joint loss function based multi-task learning and the balance coefficient in CS recognition with limited data size. Besides, the performance on the multi-speaker CS dataset shows good generalization performance for different speakers.

4. Conclusions

In this work, we proposed a cross-modal knowledge distillation method including two structures, a teacher model and a CNN-BiLSTM student model with two distillation strategies for CS recognition on limited size of datasets. We highlighted the effectiveness of our method and the importance of weighing losses based multi-task learning as well as the balance coefficient for the frame-level distillation gradient descent. The comparative evaluation demonstrated that the proposed method achieves new SOTA on the CS recognition in the multi-speaker scenario. As for future work, we will be engaged in improving the teacher and student model by language models to decrease the insertion errors.

5. Acknowledgements

This work was supported by the GuangDong Basic and Applied Basic Research Foundation (No. 2020A1515110376) and the National Natural Science Foundation of China (grant No. 61977049). The authors would like to thank the professional CS speakers from Cued Speech UK for the British English CS dataset recording.

6. References

[1] R. O. Cornett, "Cued speech," *American annals of the deaf*, pp. 3–13, 1967.

[2] P. Heracleous, D. Beautemps, and N. Aboutabit, "Cued speech automatic recognition in normal-hearing and deaf subjects," *Speech Communication*, vol. 52, no. 6, pp. 504–512, 2010.

[3] P. Heracleous, D. Beautemps, and N. Hagita, "Continuous phoneme recognition in cued speech for french," *in Proc. EUSIPCO*, pp. 2090–2093, 2012.

[4] L. Liu, T. Hueber, G. Feng, and D. Beautemps, "Visual recognition of continuous cued speech using a tandem cnn-hmm approach." *in Proc. INTERSPEECH*, pp. 2643–2647, 2018.

[5] K. Papadimitriou and G. Potamianos, "A fully convolutional sequence learning approach for cued speech recognition from videos," *in Proc. EUSIPCO*, pp. 326–330, 2021.

[6] L. Liu, G. Feng, D. Beautemps, and X.-P. Zhang, "Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 292–305, 2020.

[7] L. Liu, G. Feng, and D. Beautemps, "Automatic temporal segmentation of hand movements for hand positions recognition in french cued speech," *in Proc. ICASSP*, pp. 3061–3065, 2018.

[8] L. Liu, J. Li, G. Feng, and X.-P. S. Zhang, "Automatic detection of the temporal segmentation of hand movements in british english cued speech." *in Proc. INTERSPEECH*, pp. 2285–2289, 2019.

[9] L. Liu, G. Feng, D. Beautemps, and X.-P. Zhang, "A novel resynchronization procedure for hand-lips fusion applied to continuous french cued speech recognition," *in Proc. EUSIPCO*, pp. 1–5, 2019.

[10] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[11] N. Aboutabit, "Reconnaissance de la langue française parlée complété (lpc): décodage phonétique des gestes main-lèvres." Ph.D. dissertation, Institut National Polytechnique de Grenoble-INPG, 2007.

[12] T. Burger, A. Caplier, and S. Mancini, "Cued speech hand gestures recognition tool," *in Proc. EUSIPCO*, pp. 1–4, 2005.

[13] F. M. Thoker and J. Gall, "Cross-modal knowledge distillation for action recognition," *in Proc. ICIP*, pp. 6–10, 2019.

[14] Z. Luo, J.-T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," *in Proc. ECCV*, pp. 166–183, 2018.

[15] E. J. Crowley, G. Gray, and A. J. Storkey, "Moonshine: Distilling with cheap convolutions." *in Proc. NeurIPS*, pp. 2893–2903, 2018.

[16] W. Hao and Z. Zhang, "Spatiotemporal distilled dense-connectivity network for video action recognition," *Pattern Recognition*, vol. 92, pp. 13–24, 2019.

[17] M.-C. Wu, C.-T. Chiu, and K.-H. Wu, "Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks," *in Proc. ICASSP*, pp. 2202–2206, 2019.

[18] S. Zhang, S. Guo, L. Wang, W. Huang, and M. Scott, "Knowledge integration networks for action recognition," *in Proc. AAAI*, vol. 34, no. 07, pp. 12 862–12 869, 2020.

[19] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," *in Proc. CVPR*, pp. 2827–2836, 2016.

[20] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," *in Proc. CVPR*, pp. 7356–7365, 2018.

[21] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *in Proc. ICASSP*, pp. 6645–6649, 2013.

[22] M. Huang, Y. You, Z. Chen, Y. Qian, and K. Yu, "Knowledge distillation for sequence model." *in Proc. INTERSPEECH*, pp. 3703–3707, 2018.

[23] R. Takashima, L. Sheng, and H. Kawai, "Investigation of sequence-level knowledge distillation methods for ctc acoustic models," *in Proc. ICASSP*, pp. 6156–6160, 2019.

[24] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," *in Proc. ICML*, pp. 173–182, 2016.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *in Proc. ICML*, pp. 448–456, 2015.

[26] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *in Proc. ICML*, pp. 369–376, 2006.

[27] R. Schalkoff, *Artificial Neural Networks*. McGraw-Hill New York, 1997.

[28] C. Benoit, T. Lallouache, T. Mohamadi, and C. Abry, "A set of french visemes for visual speech synthesis," pp. 485–501, 1992.

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[30] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," *in Proc. CVPR*, pp. 7482–7491, 2018.

[31] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[32] F. Béchet, "Lia phon: un systeme complet de phonétisation de textes," *Traitement automatique des langues*, vol. 42, no. 1, pp. 47–67, 2001.