IMPROVING AUDIO-VISUAL SPEECH RECOGNITION PERFORMANCE WITH CROSS-MODAL STUDENT-TEACHER TRAINING

Wei Li[1], Sicheng Wang[1], Ming Lei[2], Sabato Marco Siniscalchi[1,3] and Chin-Hui Lee[1]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
[2]Machine Intelligence Technology, Alibaba Group, Beijing, China
[3]Department of Computer Engineering, Kore University of Enna, Enna, Italy
{lee.wei, sichengwang}@gatech.edu, lm86501@alibaba-inc.com,
marco.siniscalchi@unikore.it, chl@ece.gatech.edu

ABSTRACT

In this paper, we propose a cross-modal student-teacher learning framework to make a full use of externally abundant acoustic data in addition to a given task-specific audio-visual training database for improving speech recognition performance under the low signal-to-noise-ratio (SNR) and acoustic mismatch conditions. First, a teacher model is trained with large-sized audio-only databases. Next, a student, namely a deep neural network (DNN) model, is trained on a small-sized audio-visual database to minimize the Kullback-Leibler (KL) divergence between its output and the posterior distribution of the teacher. We evaluate the proposed approach in both matched and mismatch acoustic conditions for phone recognition with the NTCD-TIMIT database. Compared to the DNN recognition system trained with the original audio-visual data only, the proposed solution reduces the phone error rate (PER) from 26.7% to 21.3% on a matched acoustic scenario. In the mismatch conditions, the PER is reduced from 47.9% to 42.9%. Moreover, we show that posteriors generated by the teacher contain environmental information, which enables our proposed student-teacher learning to work as an environmental-aware training and good PER reductions are observed in all SNR conditions.

Index Terms— Audio-visual speech recognition, deep neural network, cross-modal training, student-teacher training, transfer learning, environmental-aware training

1. INTRODUCTION

Human machine interfaces (HMIs), such as those available in the smartphones, autonomous vehicles and robots, are emerging in our daily life. Therefore, advanced noise-robust automatic speech recognition (ASR) engines are needed for achieving effective HMI. In recent years, deep neural network (DNN) based acoustic model and its variations, e.g., long-short-term memories (LSTMs), are reported to achieve state-of-the-art ASR accuracies, e.g., [1, 2, 3]. However, ASR robustness still remains a challenge for DNN-based acoustic models. For example, degradation of the DNN performance is observed in [4] when the signal-to-noise-ratio (SNR) drops. The performance of DNNs also degrades significantly when an acoustic model trained with close-talk speech is tested on far-field recordings [5]. An effective way to increase the robustness of DNN model is to reduce the mismatch between training and testing conditions via speech enhancement [6-9]. Meanwhile i-vectors can be incorporated as speaker representation or environment information [10-11] into the original acoustic features for speaker-aware or environmental-aware DNN training.

Motivated by the fact that speech perception is a bi-modal (audio-visual) process, another line of effort for improving the robustness of ASR models is to combine cross-modal information, e.g., the spectrogram of acoustic data and lip movement contour are used to train an audio-visual ASR [12-25] in which visual information, e.g., lip movement and shape, is extracted as eigenlips or DCT features. Over the past few decades, many integration strategies have been investigated and can be divided into three major categories. In *feature fusion* [12-16], extracted visual information is concatenated with the original acoustic information (e.g., mel-frequency cepstral coefficients (MFCCs) [26]) to be subsequently used for audio-visual model training and ASR decoding. In *decision fusion* [15-17], posterior scores of DNNs trained for each modality are weighted and combined to make the final ASR prediction. In *intermediate fusion*, multi-stream hidden Markov models (HMMs) [18-22], coupled HMMs [23-24] and the turbo decoder [25] are proposed to seek the audio-visual complementarity at a HMM state level. Comprehensive comparison among the abovementioned fusion strategies can be found in [15], where decision fusion achieved the lowest phone accuracy, and intermediate fusion performed slightly better than feature fusion.

Although the previous proposed audio-visual systems have achieved satisfactory ASR performances, there is still room for further improvements. Specifically, current gains are often limited to the size of audio-visual cross-modal databases, which are far smaller than available audio-only databases, e.g., LIBRISPEECH [27] with approximately 1000 hours of read speech. In fact, most reported audio-visual ASR experiments have been conducted using commands/digits-like databases such as GRID [28] and CUAVE [29]. Due to capturing difficulties and budget constraints, recently released TCD-TIMIT [30] audio-visual database for English phone recognition has only around seven hours of parallel data. Therefore, how to use external abundant audio data to help training audio-visual models is a promising direction for further improvement. For example, in [31], audio-only GMM-HMM is first trained with an external large audio database. Then the best phone state sequence on the training set is generated by forced-alignment. Finally, a visual GMM-HMM is trained on pairs of visual feature vectors and the abovementioned state sequence. Reported gains in [31] are mainly due to the better state-level alignments for visual model training, and enhanced audio models.

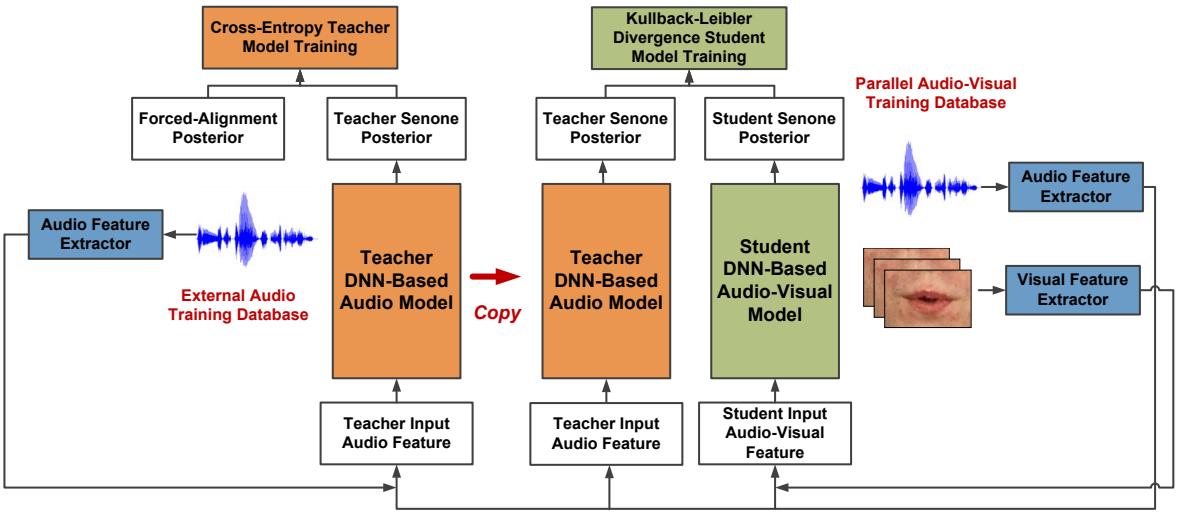


Figure 1: *Overview of the cross-modal student-teacher training framework*

In this paper, we propose a novel cross-modal student-teacher training framework to make a full use of abundant audio material. Specifically, a student audio-visual DNN is trained to minimize the Kullback-Leibler (KL) divergence between the student's output and the posteriors generated by a teacher DNN acoustic model well-trained on large audio-only databases. The generated posteriors can be viewed as teacher's knowledge learned from large audio corpora, namely each frame is labeled as a posterior vector not only containing phone information, but also embedding environmental information, e.g., noise types and SNR levels - an example will be later shown in the experiment section. In contrast to the conventional acoustic model training, where only phone labels are given, the proposed student-teacher framework is therefore environmental-aware training. Compared to the original system trained only on limited audio-visual data, our solution attains good phone error rate reductions in both matched and mismatch conditions.

2. OVERVIEW OF THE CROSS-MODAL STUDENT-TEACHER TRAINING FRAMEWORK

Figure 1 shows the proposed cross-modal student-teacher training framework, consisting of DNN-based teacher audio model training and student audio-visual model training. In the first training stage, the teacher audio model is trained by external audio data in addition to the audio portion of the given audio-visual database, and the cross-entropy criterion [1] is used. In the second training stage, a student audio-visual DNN model is learned by minimizing the KL divergence between its network output and the posterior probability generated by the trained/copied teacher, in which the input to the student model is cross-modal features and its audio part features are fed into the teacher model for posteriors generation. To focus on effect of our proposed solution, we only adopt *feature fusion* cross-modal model training in this work.

2.1. Cross-Entropy Training for Audio Teacher Model

Deep acoustic/audio models in speech applications are often trained with acoustic features and their corresponding forced-aligned labels generated from trained generative models, e.g. GMM-HMM [1]. The traditional cross-entropy training shown in Eq. (1) is used to minimize the difference between model's output and forced-aligned labels.

$$L^{(CE)}(\theta) = -\sum_{t} P_T(s_i|x_{a,t}) \quad (1)$$

where $P_T(s_i|x_{a,t})$ is the posterior generated by the *teacher* DNN, $x_{a,t}$ denotes the acoustic features at time t, and s_i is the i-th shared state (senone) [1] representing the ground true state-label at time t.

2.2. Training Audio-Visual Cross-Modal Student Model

The basic idea of traditional student-teacher training is using posteriors generated by trained teacher to guide the training of student models, e.g., compact DNN for model compression [32], enhanced DNN for knowledge distillation [33]. Teacher and student models are both trained using the audio modality information. In this paper, KL-divergence training shown in Eq. (2) is adopted to transfer the knowledge learned from extensive audio data to audio-visual DNN trained with limited cross-modal data. The transferred knowledge is expected to alleviate the shortage of audio-visual data.

$$L^{(KL)}(\theta) = \sum_{t}\sum_{i} P_T(s_i|x_{a,t}) log \frac{P_T(s_i|x_{a,t})}{P_S(s_i|x_{av,t})} \quad (2)$$

where $P_T(s_i|x_{a,t})$ and $P_S(s_i|x_{av,t})$ are the posteriors generated by teacher and student networks, $x_{a,t}$ and $x_{av,t}$ denote the acoustic feature and concatenated audio-visual feature at time t, and s_i is the i-th senone, respectively. A gradient decent method is used to update the parameter θ associated with student audio-visual DNN, while DNN parameter θ of trained/copied teacher in Eq. (1) remains fixed.

3. EXPERIMENTS

3.1. Speech Corpora

Two speech corpora, (i) NTCD-TIMIT [16], down-sampled version of recently released audio-visual database TCD-TIMIT [30], and (ii) 100-hour audio data, randomly down sampled from LIBRISPEECH [27], are mixed to train DNN-based audio and audio-visual models. In this paper, the 100-hour LIBRISPEECH subset is treated as an external audio database for training purposes.

Table 1: *Details of audio-visual corpus used in our experiment*

	Training Set	Development Set	Testing Set
Hours	≈5 h	≈1 h	≈1 h
Speakers	39	8	9
Utterances	3822	784	882

Table 1 reports additional information on the clean speech portion selected from NTCD-TIMIT. To construct the multi-condition acoustic model, each NTCD-TIMIT utterance is added with 5 noise types selected from [34] at 5 different SNR levels from 0dB to 20dB at a 5dB interval. The same data augmentation method is applied to the 100-hour LIBRISPEECH dataset, but one more down sampling step is executed to avoid a fully-expanded 2500-hour training set. Next, the abovementioned augmented datasets are merged to train the DNN-based audio teacher model, and the expanded audio-visual set is used to train the student model. Finally, a mismatched testing set is created by mixing another five different noise types from [34] with the test set in Table 1.

3.2. Audio Teacher DNN Model Training Setup

The speaker-independent acoustic model is trained with the open source Kaldi toolkit [35], and the shared scripts in [16]: a CD-GMM-HMM acoustic model is initially trained with the speaker-adaptive-training criterion. Then the CD-DNN-HMM model is built using alignments provided by the CD-GMM-HMM system. The DNN has six hidden layers each containing 1024 sigmoid units. The DNN input spans a window of 11 consecutive speech frames. Each frame contains 40-dimensional feature-space maximum likelihood linear regression (fMLLR) features.

3.3. Audio-Visual Student DNN Model Training Setup

Except for the input dimension, the speaker-independent audio-visual DNN model has the same architecture as its teacher model. Namely, 40-dimensional fMLLR visual feature extracted like in [16] is concatenated with 40-dimensional fMLLR audio feature as an audio-visual feature vector fed for DNN training. The audio-visual student model is trained to minimize KL divergence between its output and the posteriors generated by the trained teacher, where the input of student model is audio-visual feature and its audio part feature is fed into the teacher model for posteriors generation. A biphone language model is built using the training set in Table 1.

3.4. Experimental Results and Discussions

Table 2 shows the PER for different audio and audio-visual ASR systems. The baseline DNN audio system achieves a 27.3% PER, and extra visual information reduces it to 26.7%. This improvement is much smaller than reported result in many published papers [12-25], where acoustic mismatch between training and testing set is designed to show the importance of visual information. Therefore, we evaluate the first two DNN models on unmatched testing sets. The PER is reduced from 60.4% to 47.9% (as shown in Figure 2), which confirms previous argument that extra visual features can reduce PER under the adverse acoustic conditions.

We next evaluate the improvement brought by extra audio data from the LIBRISPEECH database. The DNN audio model trained with more data achieves a better result, and the PER is further reduced to 24.1%. The improvement is mainly caused by a much larger-sized training set, covering a broader acoustic space.

Table 2: *phone error rate (PER) on matched condition test sets. The first column denotes the training sets and modalities for training teacher model. The second column indicates the training set and modalities for training student model. If none, the teacher model is used for evaluation. The third column shows the PER.*

Train Teacher	Train Student	PER
audio data NTCD-TIMIT	none	27.3%
audio-visual data NTCD-TIMIT	none	26.7%
audio data LIBRISPEECH+NTCD-TIMIT	none	24.1%
audio data LIBRISPEECH+NTCD-TIMIT	audio data NTCD-TIMIT	24.4%
audio data LIBRISPEECH+NTCD-TIMIT	audio-visual data NTCD-TIMIT	22.8%

Subsequently, our proposed student-teacher training framework is compared with the abovementioned baselines. In this case, conventional student-teacher training [32-33] is first evaluated on the same modality space, namely, student and teacher models are trained using only audio features. The fourth and fifth rows in Table 2 show that student-teacher training can let a DNN, trained with only a small amount of audio data, achieve similar performance with the teacher DNN trained with the full set data. Similar to previous published results [32-33], the student's PERs should be close to, but cannot surpass the teacher's performance, because the student DNN is trained with teacher's posterior distribution using the same audio features. Next, we use cross-modal (audio-visual) features to let student mimic teacher's performance. Surprisingly, the student DNN, with 22.8% PER, beats the teacher with 24.1% PER and achieves the lowest PER. One reason is that the student model uses better/complementary features, e.g., audio + visual, compared with the teacher, where only audio information is used. At testing stage, cross-modal features can help student easily map the input features into the expected posterior distribution. In contrast, the teacher trained with only one modality features faces a bigger challenge at mapping input feature into the expected posteriors learned from the training set.

At first glance, extra visual information and cross-modal student-teacher training both bring improvement over audio-only ASR systems. To have a deeper understanding of the source of the gains, we broke down the PERs of the test set according to different SNR levels. The SNR-dependent PERs are summarized in Tables 3-7. Tables 3 and 4 shows that visual features are not always helpful in higher SNR conditions, e.g., 15dB and 20dB. This observation is consistent with previous reported results in [21, 22], where a strategy equally reliant on audio and visual information is adopted, as in this work. Obviously, we need to pay more attention to audio streams at high SNR. Complementary information from the visual modality is instead useful in low SNR conditions.

More interestingly, we find that student-teacher cross-modal training achieves lower PER in all SNR conditions than its audio-only counterpart shown in the Tables 5 and 7 although audio and visual features are still equally treated. It seems that posteriors generated by the teacher contain environmental information, making student-teacher training acting as an environmental-aware training, where DNN is trained to learn the mapping function between audio-

Table 3: SNR-dependent *PER on test set, average PER is 27.3%*

Train Teacher: audio data NTCD-TIMIT Train Student: none				
0dB	5dB	10dB	15dB	20dB
36.7%	29.9%	25.5%	22.9%	21.6%

Table 4: SNR-dependent *PER on test set, average PER is 26.7%*

Train Teacher: audio-visual data NTCD-TIMIT Train Student: none				
0dB	5dB	10dB	15dB	20dB
34.5%	28.8%	25.2%	23.1%	22.0%

Table 5: SNR-dependent *PER on test set, average PER is 24.1%*

Train Teacher: audio data LIBRISPEECH+NTCD-TIMIT Train Student: none				
0dB	5dB	10dB	15dB	20dB
33.8%	26.5%	22.1%	19.6%	18.3%

Table 6: SNR-dependent *PER on test set, average PER is 24.4%*

Train Teacher: audio data LIBRISPEECH+NTCD-TIMIT Train Student: audio data NTCD-TIMIT				
0dB	5dB	10dB	15dB	20dB
34.3%	26.9%	22.4%	19.9%	18.5%

Table 7: SNR-dependent *PER on test set, average PER is 22.8%*

Train Teacher: audio data LIBRISPEECH+NTCD-TIMIT Train Student: audio-visual data NTCD-TIMIT				
0dB	5dB	10dB	15dB	20dB
32.5%	25.1%	20.7%	18.4%	17.2%

Table 8: SNR-dependent *PER on test set, average PER is 21.3%*

Train Teacher: audio data LIBRISPEECH+NTCD-TIMIT Train Student: audio-visual data NTCD-TIMIT **Combine CE and KL Training [36] (equal weight)**				
0dB	5dB	10dB	15dB	20dB
28.5%	22.9%	19.8%	18.1%	17.2%

visual features and the corresponding labels, including the phone identity, the noisy type and the SNR information. To visualize the SNR information contained in the posteriors generated by the teacher, one utterance in the training set is first selected and mixed with a fixed noise type at 5 different SNR levels to create five utterances in order to compare the 5x5 pair-wise KL divergence. As shown in Table 9, close SNR values between the pair usually imply low KL divergence which is calculated using Eq. (3) below:

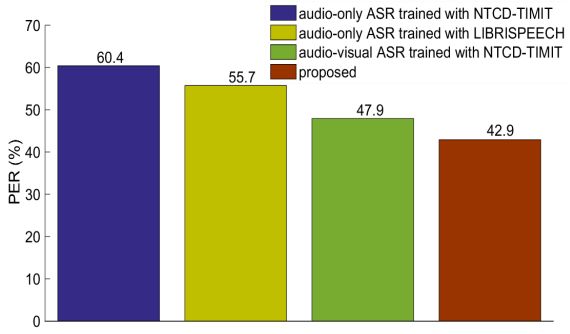
$$D_{KL}(Q||P) = \frac{1}{T}\sum_{t}\sum_{i} Q(s_i|x_{q,t}) log \frac{Q(s_i|x_{q,t})}{P(s_i|x_{p,t})} \qquad (3)$$

where $Q(s_i|x_{q,t})$ and $P(s_i|x_{p,t})$ are the frame-level posteriors generated by teacher, $x_{p,t}$ and $x_{q,t}$ denote acoustic feature at time t for utterance P and Q, and s_i is the *i*-th senone. T is utterance length.

Although our proposed cross-modal student-teacher training enhanced audio-visual ASR performance, there is still room for further improvements. Specifically, we averaged the posteriors extracted from forced-alignment and the trained teacher for better audio-visual DNN model training. In doing so, Tables 7 and 8 shows that PER is further reduced from 22.8% to 21.3%, this gain mainly comes from low SNR conditions, e.g., 0 dB and 5 dB. One reason is that posterior produced by the teacher DNN is not very accurate in lower SNRs. Therefore, combing it with force-aligned targets can make the student DNN learn more accurate senone labels. Finally, the reductions in PERs shown in Figure 2 also demonstrate our proposed system's efficiency in mismatch conditions.

Table 9: KL divergence for each utterance pair (Q and P), where utterance is represented by its corresponding SNR value.

P \ Q	0 dB	5 dB	10 dB	15 dB	20 dB
0 dB	0	0.59	1.14	1.55	1.84
5 dB	0.64	0	0.28	0.65	0.95
10 dB	1.74	0.35	0	0.15	0.37
15 dB	2.55	0.87	0.17	0	0.08
20 dB	3.06	1.28	0.45	0.09	0



4. CONCLUSIONS AND FUTURE WORK

Through a series of systematic experiments, we have shown that the performance of audio-visual ASR system for phone recognition can be improved by utilizing external audio data and cross-modal student-teacher training, where audio teacher model trained by extra large audio database can transfer learned knowledge to audio-visual student model trained with limited cross-modal data. In this paper, learned knowledge refers to posteriors generated by teacher model, it not only contains phone labels but also embeds environmental information, e.g., SNR levels and noise types. Therefore, the proposed student-teacher framework works as environment-aware training and provides more accurate labels to describe the input audio-visual features. Finally, systems trained with the posteriors extracted from forced-alignment and trained teacher achieve the best performance, namely our proposed system reduces the phone error rate from 26.7% to 21.3% on matched scenarios. In the mismatch condition, the phone error rate is reduced from 47.9% to 42.9%.

In the future, sequential student-teacher training proposed in [37] will be investigated. Moreover, the weights between posteriors from forced-alignment and trained teacher will also be analyzed in detail.

5. ACKNOWLEDGMENT

The first two authors performed this work while they were research interns at Alibaba iDST, Beijing, China, in summer 2018. The fourth author helped revising the work after re-joining the Kore University of Enna. The first author was also partially supported by a grant from the China Scholarship Council. The fourth author was partially supported by the Italian NFR AULUS project.

6. REFERENCES

[1] G. E. Dahl, D. Yu, and L. Deng et al., "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2012.

[2] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013.

[3] W. Xiong, J. Droppo, and X. Huang et al., "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[4] Y. Huang et al., "A comparative analytic study on the gaussian mixture and context dependent deep neural network hidden markov models," in *Proc. INTERSPEECH*, 2014.

[5] J. Li, R. Zhao, and Z. Chen et al., "Developing far-field speaker system via teacher-student learning," in *Proc. ICASSP*, 2018.

[6] Y. Xu, J. Du, and L. Dai et al., "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, 2014.

[7] B. Wu, K. Li, and M. Yang et al., "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[8] F. Weninger, H. Erdogan, and S. Watanabe et al., "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*, 2015.

[9] Y. Xu, J. Du, and L. Dai et al., "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015.

[10] V. Gupta, P. Kenny, and P. Ouellet et al., "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *Proc. ICASSP*, 2014.

[11] Y. Miao and F. Metze, "Distance-aware dnns for robust speech recognition," in *Proc. INTERSPEECH*, 2015.

[12] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[13] K. Thangthai, R.W. Harvey, and S.J. Cox et al., "Improving lipreading performance for robust audiovisual speech recognition using dnns," in *Proc. AVSP*, 2015.

[14] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. ICASSP*, 2015.

[15] A. H. Abdelaziz, "Comparing fusion models for dnn-based audiovisual continuous speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[16] A. H. Abdelaziz, "Ntcd-timit: A new database and baseline for noise-robust audio-visual speech recognition," in *Proc. INTERSPEECH*, 2017.

[17] P. Teissier, J. Robert-Ribes, and J.L. Schwartz et al., "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 1999.

[18] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, 2000.

[19] A. H. Abdelaziz, "Improving acoustic modeling using audio-visual speech," in *Proc. ICME*, 2017.

[20] K. Noda, Y. Yamaguchi, and K. Nakadai et al., "Audio-visual speech recognition using deep learning," *Applied Intelligence*, 2015.

[21] H. Meutzner, N. Ma, and R. Nickel et al., "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," in *Proc. ICASSP*, 2017.

[22] H. Ninomiya, N. Kitaoka, and S. Tamura et al., "Integration of deep bottleneck features for audio-visual speech recognition," in *Proc. INTERSPEECH*, 2015.

[23] A. V. Nefian, L. Liang, and X. Pi et al., "A coupled hmm for audio-visual speech recognition," in *Proc. ICASSP*, 2002.

[24] A. H. Abdelaziz, "Dynamic stream weight estimation in coupled-hmm-based audio-visual speech recognition using multilayer perceptrons," in *Proc. INTERSPEECH*, 2014.

[25] S. Gergen, S. Zeiler, and A.H. Abdelaziz et al., "Dynamic stream weighting for turbo-decoding-based audiovisual asr," in *Proc. INTERSPEECH*, 2016.

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980.

[27] V. Panayotov, G. Chen, and D. Povey et al., "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[28] M. Cooke, J. Barker, and S. Cunningham et al., "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, 2006.

[29] E. K. Patterson, S. Gurbuz, and Z. Tufekci et al., "Cuave: A new audio-visual database for multimodal human-computer interface research," in *Proc. ICASSP*, 2002.

[30] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, 2015.

[31] S. Kalantari, D. Dean, and H. Ghaemmaghami et al., "Cross database training of audio-visual hidden markov models for phone recognition," in *Proc. INTERSPEECH*, 2015.

[32] J. Li, R. Zhao, and J.T. Huang et al., "Learning small-size dnn with output-distribution-based criteria," in *Proc. INTERSPEECH*, 2014.

[33] J. Ba and R. Caruana, "Do deep nets really need to be deep?," in *Proc. NIPS*, 2014.

[34] G. Hu, "100 nonspeech sounds," in *http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html*, 2004.

[35] D. Povey, A. Ghoshal, and G. Boulianne et al., "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[36] T. Asami, R. Masumura, and Y. Yamaguchi et al., "Domain adaptation of dnn acoustic models using knowledge distillation," in *Proc. ICASSP*, 2017.

[37] J.H.M. Wong and M.J. Gales, "Sequence student-teacher training of deep neural networks," in *Proc. INTERSPEECH*, 2016.