End-to-end Whispered Speech Recognition with Frequency-weighted Approaches and Layer-wise Transfer Learning

Heng-Jui Chang, Alexander H. Liu, Hung-yi Lee, Lin-shan Lee

College of Electrical Engineering and Computer Science, National Taiwan University

{b06901020, r07922013, hungyilee, lslee}@ntu.edu.tw

Abstract

Whispering is an important mode of human speech, but no end-to-end recognition results for it were reported yet, probably due to the scarcity of available whispered speech data. In this paper, we present several approaches for end-to-end (E2E) recognition of whispered speech considering the special characteristics of whispered speech and the scarcity of data. This includes a frequency-weighted SpecAugment policy and a frequency-divided CNN feature extractor for better capturing the high frequency structures of whispered speech, and a layer-wise transfer learning approach to pre-train a model with normal speech then fine-tuning it with whispered speech to bridge the gap between whispered and normal speech. We achieve an overall relative reduction of 19.8% in PER and 31.9% in CER on a relatively small whispered TIMIT corpus. The results indicate as long as we have a good E2E model pre-trained on normal speech, a relatively small set of whispered speech may suffice to obtain a reasonably good E2E whispered speech recognizer.

Index Terms: whispered speech, end-to-end speech recognition, data augmentation, transfer learning

1. Introduction

Although less frequently used than normal speech, whispering is an important mode of human speech used in special occasions such as interchanging confidential information, having conversations in meetings, theaters, libraries or bedrooms, or for patients with impaired glottises. Machine recognition of whispered speech is crucial yet extremely difficult due to the very special nature of whispered speech, such as no vocal cord vibrations [1, 2], lower speaking rates [1, 3], lower energy [4, 5, 6], upward shift of formant frequencies [4, 5], flatter spectra [4, 5, 6, 7], etc. Automatic speech recognition (ASR) systems trained on normal speech thus inevitably degraded severely for whispered speech due to such mismatch [5, 6]. Various approaches have been used to overcome these difficulties. Good examples included model adaptation [2, 8, 9], pseudo whisper features [5, 6, 10], non-audible murmur microphone (NAM) [11], articulatory features [8, 12, 13], and visual cues [14, 15, 16], achieving substantial improvements primarily based on the earlier very successful hidden Markov models (HMM) [17].

Recently, E2E ASR approaches such as connectionist temporal classification (CTC) [18], RNN-transducer [19], and Sequence-to-sequence model [20] have been overwhelmingly attractive and shown to be effective in globally optimizing the whole ASR process for the overall performance rather than locally optimizing acoustic and language models under different criteria. These approaches achieved very exciting accuracy but required only training data, without need for hand-crafted modules or language-specific knowledge as in earlier approaches.

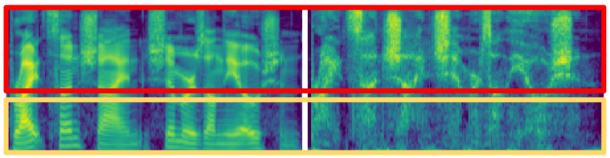


Figure 1: *Mel-spectrograms of the same sentence produced by the same speaker in normal and whispered voice. The high frequency features (red box) are preserved in whispered speech, while the low frequency features (yellow box) are seriously lost.*

However, the effectiveness of E2E approaches over whispered speech is yet to be confirmed. Previous works suggested that deep learning was useful for whispered speech recognition [21, 22, 23], while the success of E2E approaches on normal speech ASR was widely believed to depend on the quantity of data [24] and the model architecture [25]. It is much more difficult to collect whispered speech data of reasonable size, and the special characteristics of whispered speech may need special considerations in model design and training. These are the questions this paper wishes to obtain at least some answers to.

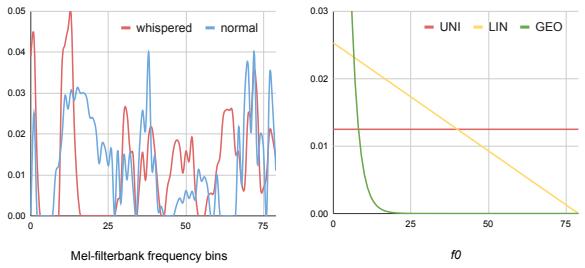
This paper is to our knowledge the earliest report focusing on whispered speech recognition with E2E models. We propose a frequency-weighted SpecAugment [26] policy, a frequency-divided CNN extractor, and a layer-wise transfer learning approach to bridge the gap between whispered and normal speech. We achieve an overall relative reduction of 19.8% in PER and 31.9% in CER on a relatively small whispered TIMIT corpus [2], which is already a very narrow gap from the performance for normal speech.

2. Proposed Methods

This work is based on the CTC model [18] for E2E ASR consisting of a deep CNN feature extractor [27] and a multi-layer bidirectional LSTM. The model takes a sequence of acoustic features $\boldsymbol{x} = (x_1, \ldots, x_T)$ with length T for the input utterance. It is encoded first by the deep CNN extractor performing downsampling, and further by the BLSTMs to obtain a sequence of hidden states. This sequence is then linearly transformed into a sequence $\boldsymbol{y} = (y_1, \ldots, y_{T'})$, where each y_t represents a probability distribution over all possible output symbols at each time index, and $T' < T$. The ASR model is trained to minimize the CTC loss function [18, 28].

2.1. Analysis for Frequency Importance

It has been well known that the characteristics for normal speech are reasonably preserved in whispers for higher frequencies, but seriously lost in lower frequencies, as shown in an example in Fig. 1 and [1, 2, 4, 5, 6, 7]. So, we suspect the



(a) (b)

Figure 2: *(a) The weight distributions $\hat{\mathbf{w}}$ obtained in the experiment described in Sec. 2.1 for whispered (red) and normal (blue) speech, (b) the uniform (UNI), linearly (LIN) and geometrically (GEO) decreasing distributions for sampling the lower end f_0 of the mask in SpecAugment.*

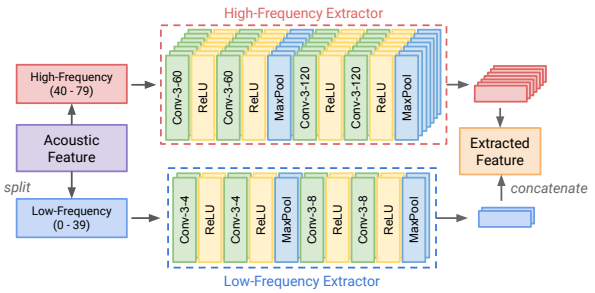


Figure 3: *The frequency-divided CNN extractor. Conv-k-c denotes 2D convolution with kernel size of k and c output channels. The low-frequency extractor has fewer convolutional filters in order to compress the features.*

higher frequencies are more important to E2E ASR for whispered speech, although both high and low frequencies play important roles for normal speech. We first analyze this assumption here.

We use two E2E ASR systems pre-trained with normal and whispered speech respectively for the experiment below. We define a learnable weight vector $\mathbf{w} = [w_0\ w_1\ \ldots\ w_{\nu-1}]^T$ for all the Mel-frequency bins, where ν is the total number of the frequency bins of the considered Mel-spectrogram. This vector $\mathbf{w}$ is first transformed into a probability distribution by softmax, $\hat{\mathbf{w}} = [\hat{w}_0\ \hat{w}_1\ \ldots\ \hat{w}_{\nu-1}]^T = \mathrm{softmax}(\mathbf{w})$, then used to weight the respective Mel-filterbank features,

$$x'_{t,f} = x_{t,f} \cdot \exp(-\hat{w}_f / r), \tag{1}$$

where $x_{t,f}$ is the feature for the f^{th} Mel-frequency bin at time t, $x'_{t,f}$ is the weighted value, and r is a positive scaling factor. So the features are suppressed more if the corresponding weights are higher. The weighted features are then fed to the pre-trained E2E ASR systems to learn the weight distribution $\hat{\mathbf{w}}$ for maximizing the CTC loss function, which is supposed to be minimized. So those frequency bins suppressed more by higher weights are those more important for ASR. Stochastic gradient ascent is used to obtain the learnable weight $\hat{\mathbf{w}}$.

With the experimental setup to be described in Sec. 3.1 below, the results for the weight distribution $\hat{\mathbf{w}}$ are in Fig. 2a. We see the learned weights were very different for E2E ASR for whispered and normal speech. For whispered speech more emphasis was clearly on higher frequencies, indicating more important information was there. Although some low frequencies were also weighted highly, these frequencies were actually equally important for both whispered and normal speech. On the other hand, for normal speech the E2E ASR turned out to pay almost the same level of attention to either low or high frequencies. This led to the approaches proposed below.

2.2. Frequency-weighted SpecAugment

Frequency masking of SpecAugment [26] has been shown to be an effective method for data augmentation for E2E ASR models. It is summerized below. A mask size of Δf is first sampled from a frequency range $[F_1, F_2]$ uniformly. The lower end of the mask, f_0, is then sampled uniformly from $[0, \nu - \Delta f)$, where ν is the total number of frequency bins of the spectrogram. This defines the mask $[f_0, f_0 + \Delta f)$, in which all frequency bins are set to zero when masked.

With the observation in Fig. 2a, we try to mask lower frequencies more often for whispered speech. This is referred to as *Frequency-weighted SpecAugment*, in which instead of sampling f_0 uniformly from $[0, \nu - \Delta f)$ as mentioned above, it can be sampled from a linearly or geometrically decreasing distribution as shown in Fig. 2b. The probability that the lower frequency bins being masked would thus be higher, or the machine would learn less precise information from, or rely less on lower frequencies.

2.3. Frequency-divided CNN Extractor

Since the standard deep CNN extractor [27] used for E2E ASR treats all frequencies equally, here we propose a *Freqency-divided CNN extractor* containing two CNN extractors respectively processing the lower and higher frequency half of the features separately as shown in Fig. 3. Though with the same total number of feature parameters as the standard extractor, the low-frequency extractor has fewer filters. Therefore, the high-frequency extractor with more filters can capture more cues and offer more information from the preserved structures in high frequency regions of whispered speech.

2.4. Layer-wise Transfer Learning from Normal Speech

The scarcity of whispered speech data makes training E2E ASR challenging, but much more data for normal speech are available. We therefore propose to perform transfer learning [29] by having an E2E ASR model pre-trained on a large normal speech corpus fine-tuned on a smaller whispered speech corpus. Because BLSTMs are prone to overfit [30], fine-tuning the whole model did not work well. But since the goal is to transfer between speech types with differences primarily in acoustic characteristics, we propose to fine-tune only the bottom layers closer to the acoustic features. This layer-wise transfer learning is similar to but different from that reported for transfer between different languages, in which fine-tuning top layers allowed better transfer [29].

3. Experiments

3.1. Experimental Setup

The following two datasets were used in our experiments:
wTIMIT. The whispered TIMIT corpus [2] consisted of parallel whispered and normal speech data each around 26 hours, including 48 speakers whispering and speaking 450 phonetically balanced sentences chosen from TIMIT [31]. It was originally partitioned into the TRAIN and TEST set randomly. However, the existence of many overlapping utterances with the same sentences spoken by different speakers made it difficult to estimate

Table 1: *PERs(%) on whispered for hybrid and E2E models trained on wTIMIT whispered speech with different corpus partitions.*

Corpus Partition	ASR Model	Whispered Dev	Whispered Test
(I) Original	(a) HMM-Hybrid	35.0	35.5
	(b) E2E	13.0	13.7
(II) Ours (400/25/25)	(c) HMM-Hybrid	39.6	38.7
	(d) E2E	35.6	35.9

the actual recognition accuracy, as will be shown later in Table 1. We thus re-partitioned the dataset into train/dev/test sets, each containing 400/25/25 sentences split from the 450 sentences.

LibriSpeech. The LibriSpeech English corpus [32] included approximately 1000 hours of speech. We used the 460-hour clean set to be the additional large corpus for normal speech for transfer learning.

For comparing with HMM-based ASR, a DNN-HMM hybrid system [33] baseline was constructed using the TIMIT recipe *nnet2* from the Kaldi toolkit [34]. 13-dimensional MFCC features with delta and delta-delta were used as the recipe did. For E2E ASR, 80-dimensional log Mel-filterbank features with delta and normalization were used. Two E2E ASR models were used. The standard model used a 4-layered BLSTM of 512 units per direction and a CNN feature extractor [27]. Another light model with a 3-layered bidirectional GRU with 128 units per direction was used for small training sets such as the experiment producing the results in Fig. 2a to prevent overfitting. All results reported below were averaged over 3 runs.

3.2. E2E v.s. Hybrid for Whispered Trained on Whispered

We first compared the HMM-based hybrid model with standard E2E ASR without any approach proposed here, assuming only the whispered part of wTIMIT was available for training. The phoneme level (total 39 phonemes) annotation was used to train both models from scratch, and the phoneme error rates (PER) are in Table 1. The light E2E ASR model used to produce Fig. 2a was used here.

Section (I) of Table 1 is for the original TRAIN/TEST partition provided by wTIMIT, in which we see the E2E ASR model offered a very low error rate (row(b)) as a result of the overlapping utterances between the TRAIN/TEST sets. So all experiments below were based on our partition as mentioned in Sec. 3.1, with results in Section (II). Here we see the E2E model was slightly better than the hybrid (rows(d) v.s. (c)), even with only 26 hours of training data, for which hybrid typically outperformed E2E. This indicated E2E ASR was a more proper choice for whispered speech if the data set was not too small.

3.3. Proposed Frequency-weighted Approaches with Small Normal Speech Training

Now, we considered the case when only limited normal speech (i.e. the 26 hours of normal speech from wTIMIT) was available for training E2E ASR, but with the several approaches proposed here, to verify these approaches were useful for whispered speech regardless of the training data. The light model same as that used in Sec. 3.2 was used. The results are listed in Table 2.

Trained with Limited Normal Speech Section (I) of Table 2 is for the baselines with the zero whispered speech resource scenario without any approach proposed here. The fact that the model performance degraded seriously for whispered speech (columns(B) v.s. (A)) and E2E performed better on normal speech yet worse on whispered (rows(b) v.s. (a)) aligned with the mismatch between whispered and normal, and the assumption that E2E ASR was prone to overfit its training data [30].

Table 2: *PERs(%) on wTIMIT with only a small normal speech set for training. Section (I) for baselines, Section (II) with the proposed frequency-weighted SpecAugment (FreqSpecAug) with a uniform (UNI), linearly (LIN) and geometrically (GEO) decreasing distribution, Section (III) with frequency-divided CNN extractor (FreqCNN) applied in addition.*

Method	(A) Normal Dev	(A) Normal Test	(B) Whispered Dev	(B) Whispered Test
(I) **Baselines**				
(a) HMM-Hybrid	34.5	33.5	55.8	54.6
(b) E2E	**29.9**	**29.7**	60.7	59.5
(II) **E2E + FreqSpecAug**				
(c) UNI [26]	31.0	31.1	54.8	53.9
(d) LIN	30.6	30.6	52.9	51.8
(e) GEO	33.1	32.8	49.2	48.3
(III) **E2E + FreqCNN + FreqSpecAug**				
(f) UNI	33.5	32.8	52.0	51.3
(g) GEO	35.5	35.1	**48.3**	**47.7**

Frequency-weighted SpecAugment To find out the extra robustness achievable by the proposed frequency-weighted SpecAugment, we let the lower end f_0 of the mask in SpecAugment to be sampled from a uniform (UNI), linearly (LIN) or geometrically (GEO) decreasing distribution as described in Sec. 2.2, the results are in rows (c)(d)(e) of Section (II) in Table 2. We see with GEO proposed, a relative 18.8% PER reduction with respect to the baseline E2E (rows(e) v.s. (b)) and a 10.4% relative improvement compared to the original SpecAugment or UNI (rows(e) v.s. (c)) was achieved. This implied with the lower frequencies emphasized in SpecAugment, or letting E2E ASR learn less from lower frequencies, the performance on whispered speech was improved. Obviously in this way we forced the model to distill more details from higher frequencies where whispered speech is more similar to normal speech.

Frequency-divided CNN Extractor In Section (III) of Table 2 we added the frequency-divided CNN extractor as mentioned in Sec. 2.3 onto the models in Section (II). The results in (rows(f)(g)) show that the frequency-divided CNN made further PER reduction on whispered speech in addition (rows(g) v.s. (e) and (f) v.s. (c)). This verified extracting less lower frequency information with fewer filters while more fine structures or high frequency information with more filters did help, although the accuracy for normal speech was inevitably degraded. The overall relative improvement achieved by the frequency-weighted SpecAugment plus frequency-divided CNN extractor was 19.8% (rows(g) v.s. (b)), which was the setting for the experiments reported below.

3.4. Training with Extra Normal Speech Data

Here, we tried to reduce the performance gap between whispered and normal speech recognition by utilizing an additional large normal speech training corpus (LibriSpeech). The models below were trained on grapheme level (characters without lexicon) following previous works [35, 36, 37], with frequency-weighted SpecAugment and frequency-divided CNN extractor applied.

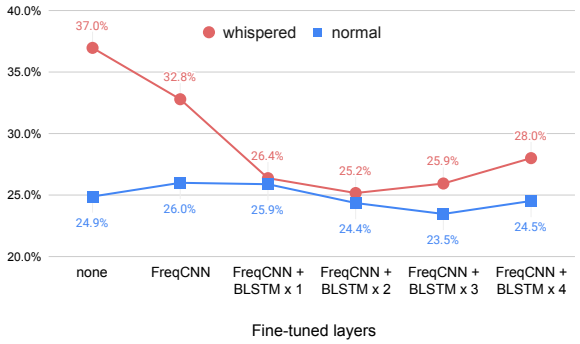


Figure 4: *CERs for whispered and normal on wTIMIT when fine-tuned from the bottom in the layer-wise transfer learning.*

Layer-wise Transfer Learning Here we studied the layer-wise transfer learning mentioned in Section 2.4. We used the 460-hr LibriSpeech normal speech data to pre-train an E2E ASR model, and then fine-tuned it with the small whispered set in wTIMIT. Instead of fine-tuning the whole model, only several bottom layers were fine-tuned. Fig. 4 depicts the results for whispered and normal speech from left to right when fine-tuning was performed on different number of bottom layers, starting with no fine-tuning.

We can see from Fig. 4 the character error rate (CER) for whispered speech was improved from 37.0% all the way to 25.2% when fine-tuning the frequency-divided CNN extractor and the first two BLSTM layers simultaneously, which was a 31.9% error rate reduction relative to the pre-trained model. Fine-tuning the 3rd BLSTM layer or further did not boost the performance, probably because layers close to the output were more related to characters and language modeling [29], and fine-tuning too many parameters affected the ability of the model and further overfitted it on the small wTIMIT corpus. We actually tried to also fine-tune the output layer, but that slightly damaged the performance probably because we were recognizing the same language. The best result of 25.2% here was only 1.9% absolutely higher than the best performance on normal speech when fine-tuning an extra layer. These results verified that the BLSTMs also played important roles in encoding acoustic features, and thus fine-tuning part of the bottom layers of a pre-trained model was helpful.

Different Methods for Training with Both Speech Types Here we wish to explore different methods using both whispered and normal speech to train the E2E ASR for whispered speech. We first set three baseline models trained solely on the wTIMIT whispered set (wTM-w), the wTIMIT normal set (wTM-n), and the LibriSpeech corpus separately, respectively in rows (a)(b)(c) in the first section of Table 3. We then used all the three sets of whispered and normal speech jointly to train the E2E ASR in rows (d)(e)(f) in the 2nd half of Table 3. This included directly sampling them randomly regardless of the size of the corpus (row(d)), oversampling whispered speech to the same size as normal speech (referred to as imbalanced learning, previously used for whisper detection [38]) (row(e)), and the layer-wise transfer learning described above (row(f)). All results in Table 3 are CERs.

First of all, the baseline models using the wTIMIT dataset performed poorly compared to using LibriSpeech (rows(a)(b) v.s. (c)). This confirmed that E2E models required a large amount of training data to work well [25]. Next, mixing all whispered and normal speech data together improved the performance slightly on the whispered set but degraded it on the normal set compared to the model using only normal speech (rows(d) v.s. (c)). Though with a relatively small amount of whispered speech (only about 5%), the E2E ASR model was still able to learn to recognize whispered speech. On the other hand, the imbalanced learning method damaged the E2E ASR severely (row(e)), probably due to the low diversity of the sentences in wTIMIT, the system thus failed to model the characters and words.

Table 3: *CERs(%) on wTIMIT when an additional normal corpus is available. wTM-w and wTM-n (rows(a)(b)) denote whispered and normal data from wTIMIT, respectively. Libri (row(c)) denotes the LibriSpeech 460-hour set as the additional data. Imbalanced learning (row(e)) is the previously used method [38]. Layer-wise TL (row(f)) denotes the layer-wise transfer learning proposed here.*

Method		Training Data	Normal		Whispered	
			Dev	Test	Dev	Test
(a)	E2E baselines (single dataset)	wTM-w	54.4	53.1	48.0	46.1
(b)		wTM-n	41.9	40.5	54.8	53.3
(c)		Libri	26.4	24.9	37.8	37.0
(d) Random Sampling		wTM-wn + Libri	28.7	28.1	34.0	32.9
(e) Imbalanced learning			43.6	41.7	47.4	45.2
(f) Layer-wise TL			**24.4**	**23.5**	**26.5**	**25.2**

In contrast, for the layer-wise transfer learning method (row(f)), we divided the training phase into two, pre-training with normal speech and fine-tuning with whispered speech. This method outperformed all other methods. Based on the model well-initialized with a large normal set, fine-tuning a part of its layers properly adapted it to whispered speech while preserving its original capability to recognize the various words in the vocabulary. In other words, the layer-wise transfer learning proposed here enables us to bridge the gap between recognizing normal and whispered speech. This implies we can use any E2E model we already have pre-trained on normal speech and do not need to collect a large amount of whispered speech.

4. Conclusions

This is the first paper focusing on exploring the possibility of E2E recognition for whispered speech. We propose a frequency-weighted SpecAugment approach and a frequency-divided CNN extractor to boost the recognition performance. With the aid of a larger normal speech corpus and a layer-wise transfer learning approach, we further show the performance gap between whispered and normal speech recognition can be reduced to very narrow.

5. References

[1] S. T. Jovičić and Z. Šarić, "Acoustic analysis of consonants in whispered speech," *Journal of Voice*, vol. 22, 2008.

[2] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, Universityof Illinois at Urbana Champaign, 2010.

[3] P. X. Lee, D. Wee, H. S. Y. Toh, B. P. Lim, N. Chen, and B. Ma, "A whispered mandarin corpus for speech technology applications," in *INTERSPEECH*, 2014.

[4] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, 2005.

[5] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Generative modeling of pseudo-whisper for robust whispered speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, 2016.

[6] Đ. T. Grozdić and S. T. Jovičić, "Whispered speech recognition using deep denoising autoencoder and inverse filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, 2017.

[7] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, 2002.

[8] Szu-Chen Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *ICASSP*, 2005.

[9] A. Mathur, S. Reddy, and R. Hegde, "Significance of parametric spectral ratio methods in detection and recognition of whispered speech," in *EURASIP Journal on Advances in Signal Processing*, 2012.

[10] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Deep neural network training for whispered speech recognition using small databases and generative model sampling," *J Speech Technol*, vol. 20, 2017.

[11] C. Yang, G. Brown, L. Lu, J. Yamagishi, and S. King, "Noise-robust whispered speech recognition using a non-audible-murmur microphone with vts compensation," in *ISCSLP*, 2012.

[12] G. Srinivasan, A. Illa, and P. K. Ghosh, "A study on robustness of articulatory features for automatic speech recognition of neutral and whispered speech," in *ICASSP*, 2019.

[13] B. Cao, M. Kim, T. Mau, and J. Wang, "Recognizing whispered speech produced by an individual with surgically reconstructed larynx using articulatory movement data," in *SLPAT*, 2016.

[14] F. Tao and C. Busso, "Lipreading approach for isolated digits recognition under whisper and neutral speech," in *INTERSPEECH*, 2014.

[15] T. Tran, S. Mariooryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in *ICASSP*, 2013.

[16] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *ICASSP*, 2018.

[17] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, 1989.

[18] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML*, 2014.

[19] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML Workshop on Representation Learning*, 2012.

[20] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[21] B. Marković, S. T. Jovičić, J. Galić, and Đ. T. Grozdić, "Whispered speech database: design, processing and application," in *Text, Speech, and Dialogue*, 2013.

[22] Đ. T. Grozdić, B. Marković, J. Galić, and S. T. Jovičić, "Application of neural networks in whispered speech recognition," *Telfor Journal*, vol. 5, 2013.

[23] B. P. Lim, F. Wong, Y. Li, and J. W. Bay, "Transfer learning with bottleneck feature networks for whispered speech recognition," in *INTERSPEECH*, 2016.

[24] C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *ICASSP*, 2018.

[25] K. J. Han, R. Prieto, and T. Ma, "State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions," in *ASRU*, 2019.

[26] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *INTERSPEECH*, 2019.

[27] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," in *INTERSPEECH*, 2017.

[28] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[29] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, "Transfer learning for speech recognition on a budget," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*. ACL, 2017.

[30] T.-S. Nguyen, S. Stüeker, J. Niehues, and A. Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," in *ICASSP*, 2020.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1992.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015.

[33] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, 2012.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[35] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *CoRR*, vol. abs/1609.03193, 2016.

[36] V. Liptchinsky, G. Synnaeve, and R. Collobert, "Letter-based speech recognition with gated convnets," *CoRR*, vol. abs/1712.09444, 2017.

[37] T. Likhomanenko, G. Synnaeve, and R. Collobert, "Who needs words? lexicon-free speech recognition," *CoRR*, vol. abs/1904.04479, 2019.

[38] T. Ashihara, Y. Shinohara, H. Sato, T. Moriya, K. Matsui, T. Fukutomi, Y. Yamaguchi, and Y. Aono, "Neural whispered speech detection with imbalanced learning," in *INTERSPEECH*, 2019.