SUBBAND TEMPORAL ENVELOPE FEATURES AND DATA AUGMENTATION FOR END-TO-END RECOGNITION OF DISTANT CONVERSATIONAL SPEECH

Cong-Thanh Do

Toshiba's Cambridge Research Laboratory, Cambridge, United Kingdom

cong-thanh.do@crl.toshiba.co.uk

ABSTRACT

This paper investigates the use of subband temporal envelope (STE) features and speed perturbation based data augmentation in end-to-end recognition of distant conversational speech in everyday home environments. STE features track energy peaks in perceptual frequency bands which reflect the resonant properties of the vocal tract. Data augmentation is performed by adding more training data obtained after modifying the speed of the original training data. Experiments show that using STE features and speed perturbation based data augmentation helps improving the performance of end-to-end speech recognition on a challenging corpus which was used for the CHiME 2018 speech separation and recognition challenge. STE features provide up to 2.0% relative word error rate (WER) reduction compared to the conventional log-Mel filter-bank (FBANK) features. Data augmentation is used with both features and provides up to 5.2% relative WER reduction. We propose a simple hypothesis selection method to combine the hypotheses produced by the end-to-end systems using FBANK and STE features. This method additionally provides up to 4.7% relative WER reduction.

Index Terms— End-to-end speech recognition, subband temporal envelope features, speed perturbation based data augmentation, CHiME 2018, distant conversational speech

1. INTRODUCTION

End-to-end automatic speech recognition (ASR) aims at using a single neural network architecture within a deep learning framework to perform speech-to-text task [1]. The architectures used for end-to-end ASR could be either attention-based encoder-decoder [2], connectionist temporal classification (CTC) [1] or hybrid CTC/attention [3]. The development of end-to-end ASR system is simplified because the training does not need a pronunciation lexicon and the whole explicit modeling of phones [1].

End-to-end ASR often use log-Mel filter-bank (FBANK) features created by skipping the discrete cosine transform (DCT) in the Mel frequency cepstral coefficient (MFCCs) computation [4]. FBANK features are spectral features which can be computed from a time-frequency representation of speech obtained with the discrete Fourier transform (DFT). The FBANK feature vectors are computed from speech frames which are independently extracted every 10 ms.

Temporal envelope features could carry temporal context information which is not explicitly extracted by the conventional spectral FBANK features. Various temporal features have been developed for ASR [5, 6, 7, 8]. In this work, we investigate the use of the subband temporal envelope (STE) features [7] in attention-based or hybrid CTC/attention end-to-end ASR in which a recurrent neural network (RNN) encoder is used. STE features track energy peaks in perceptual frequency bands which reflect the resonant properties of the vocal tract. These are temporal information about transients that are not present in the conventional spectral FBANK features. The temporal context information carried by the STE features could provide additional benefit for the RNN encoder which is better at finding and exploiting long range context from the input features [1].

For experiments, we use a large-scale corpus (CHiME-5) of real multi-speaker conversational speech recorded via commercially available multi-microphone hardware in multiple homes. This corpus was first recorded to use in the CHiME 2018 speech separation and recognition challenge [9]. The main difficulty of this corpus comes from the source and microphone distance in addition to the spontaneous and overlapped nature of the speech. In end-to-end ASR on such a challenging task, the conventional FBANK features are still widely used thanks to their effectiveness over alternative input features such as raw speech waveform [10, 11, 12, 13].

End-to-end ASR would benefit from having more data for training the neural network architecture [14, 15]. In this respect, we investigate the use of speed perturbation based data augmentation [16] for end-to-end ASR on this corpus. The performance of the STE features is evaluated in comparison with the conventional FBANK features. We show the effectiveness of the STE features and data augmentation in this challenging task for end-to-end ASR.

Combining systems using STE and FBANK features could provide additional WER reduction [7]. As the training of hybrid CTC/attention end-to-end ASR systems does not use alignment information, these systems are not able to provide outputs with exact monotonic time-aligned information [3, 17]. System combination techniques, e.g. lattice combination based on Bayes risk minimization [18] or ROVER (recognizer output voting error reduction) [19], use time-aligned information in the ASR outputs produced during the decoding to combine the ASR outputs. Without time-aligned information, it is not straightforward to apply these techniques to combine attention-based end-to-end ASR systems. In this paper, we propose a simple hypothesis selection method to combine the hypotheses produced by the systems using FBANK and STE features. This method is able to provide additional WER reduction.

The paper is organized as follows. Section 2 presents related works. The STE features extraction is presented in section 3. CHiME-5 corpus is presented in section 4. Section 5 provides details on the data augmentation method. Information on the end-to-end ASR system, namely the front-end processing, the end-to-end architecture and the system combination method, is presented in section 6. Section 7 presents the experimental results and section 8 concludes the paper.

2. RELATED WORKS

Recent works on features extraction for end-to-end ASR focus on using raw speech waveform directly as input features [10, 11, 12,

13]. In the approach using raw speech waveform as input features, the end-to-end ASR architectures are often augmented with addition layers for features computation. The computation cost is thus high and in many cases, the performance of end-to-end ASR using raw speech waveform as features is still worst than that of the system using FBANK features [1]. In [12, 13], the authors reported the first times end-to-end models trained from the raw signal significantly outperform FBANK features on Wall Street Journal (WSJ) task [20]. WSJ corpus consists primarily of read speech. The data in WSJ corpus is also quite clean and the speakers are close to microphones.

3. SUBBAND TEMPORAL ENVELOPE FEATURES

The algorithm for extracting the STE features is depicted in Fig. 1. Given a speech signal $s[n]$, the M STE signals $e_m[n]$, $m = 1, \ldots, M$ of $s[n]$ are extracted as follows. The speech signal $s[n]$ is first pre-emphasized by using a filter having a transfer function $H[z] = 1 - 0.97z^{-1}$. The pre-emphasized speech signal is then decomposed into M subband signals $s_m[n]$, $k = 1, \ldots, M$ using a filter-bank consisting of M Gammatone band-pass filters. In this work, the Gammatone filters implementation from [21] is used. Each Gammatone band-pass filter in the filter-bank is implemented as a cascade of four separate second order IIR (infinite impulse response) filters. This implementation is done to avoid round-off errors [21]. The center frequencies of the Gammatone filters are linearly spaced on the ERB (equivalent rectangular bandwidth) scale with the first one starts at 100 Hz.

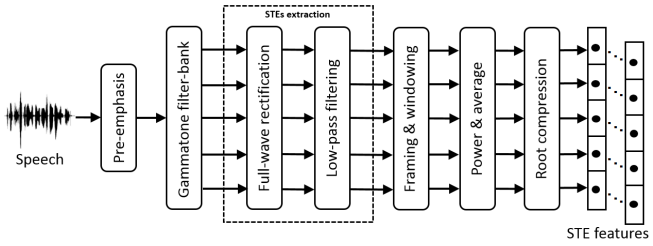

Fig. 1. Algorithm for extracting the STE features.

The STEs $e_m[n]$, $m = 1, \ldots, M$ of the subband signals $s_m[n]$, $m = 1, \ldots, M$ are then extracted by, first, full-wave rectifying the subband signals followed by a zero-phase low-pass filtering of the resulting signals. In this work, the low-pass filter for extracting STEs is a fourth-order elliptic low-pass filter with 2-dB of peak-to-peak ripple and a minimum stop-band attenuation of 50-dB. The cut-off frequency of this low-pass filter, which controls the bandwidth of the STEs, is 50 Hz because this ensures a reasonable STE bandwidth for human and machine speech recognition [22, 7]. The zero-phase filtering is performed by processing the input data in both the forward and reverse directions. After the data is filtered in the forward direction, the filtered sequence is reversed and run back through the filter. An example of the slowly-varying STE, extracted by this method, is shown in Fig. 2.

From the STEs $e_m[n]$, $m = 1, \ldots, M$ extracted from the whole utterance, short-term frames of 25 ms are extracted every 10 ms. The short-term frames are multiplied with Hamming windows to emphasize the samples in the middle of the analysis frames. At time instant k, assume that $\hat{e}_{m,k}[n], m = 1, \ldots, M$ are the short-term STEs obtained after the Hamming windowing, a feature vector $\mathbf{y}_k = [y_{1,k}, y_{2,k}, \ldots, y_{M,k}]^T$ is extracted. A feature coefficient $y_{m,k}$ is computed as follows:

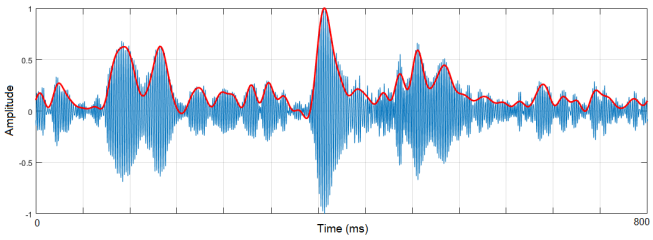

Fig. 2. Slowly-varying STE (red curve) extracted from the 10th subband signal of a speech segment of 800 ms.

$$y_{m,k} = \frac{1}{N} \sum_{n=1}^{N} \hat{e}_{m,k}^2[n],$$

where N is the number of samples in a frame and k is the frame index. The STE feature vector $\widehat{\mathbf{y}}_k$ is computed by applying the 15^{th} root on $\mathbf{y}_k$: $\widehat{\mathbf{y}}_k = [y_{1,k}^{1/15}, y_{2,k}^{1/15}, \ldots, y_{M,k}^{1/15}]^T$, according to a compression suggested in [6].

In the STE features extraction, as the STEs are first extracted from the whole utterances then the feature vectors are computed from these long-term envelopes every 10 ms, the feature coefficients can be considered as a downsampling of the temporal envelopes in the perceptual frequency bands. The variation of the feature coefficients over time reflects, to some extent, the shapes of the STEs in the perceptual frequency bands which carry important temporal cues for human speech recognition [22]. The shape of the STEs could also preserve additional temporal context information.

4. CHIME-5 CORPUS

4.1. Recording scenario

CHiME-5 is the first large-scale corpus of real multi-speaker conversational speech recorded via commercially available multi-microphone hardware in multiple homes [9]. Natural conversational speech from a dinner party of 4 participants was recorded for transcription. Each party was recorded with 6 distant Microsoft Kinect microphone arrays and 4 binaural microphone pairs worn by the participants. There are in total 20 different parties (sessions) recorded in 20 real homes. This corpus was designed for the CHiME 2018 speech separation and recognition challenge [9].

Each party has a minimum duration of 2 hours which composes of three phases, each corresponding to a different location: i) kitchen - preparing the meal in the kitchen area; ii) dining - eating meal in the dining area; iii) living - a post-dinner period in a separate living room area. The participants can move naturally within the home in different locations, but they should stay in each location for at least 30 minutes. There is no constraint on the topic of the conversations. The conversational speech is thus spontaneous.

4.2. Audio and transcriptions

The audio of the parties was recorded with a set of six Microsoft Kinect devices which were strategically placed to capture each conversation by at least two devices in each location. Each Kinect device has a linear array of 4 sample-synchronized microphones and a camera. The audio was also recorded with the Soundman OKM II Classic Studio binaural microphones worn by each participant [9].

Manual transcriptions were produced for all the recorded audio. The start and end times and the word sequences of an utterance produced by a speaker are manually obtained by listening to

the speaker's binaural recording. These information are used for the same utterance recorded by other recording devices but the start and end times are shifted by an amount that compensates for the asynchonization between devices.

4.3. Data for training and test

Training, development and evaluation sets are created from the 20 parties. Data recorded from 16 parties are used for training. The data used for training ASR systems combines both left and right channels of the binaural microphone data and a subset of all Kinect microphone data from 16 parties. In this paper, the total amount of speech used in the training set is around 167 hours (the `data/train_worn_u200k` set [9]). Each of the development and evaluation sets is created from 2 parties of around 4.5 and 5.2 hours of speech, respectively. The speakers in the training, development and evaluation sets are not overlapped.

For the development and evaluation data, information about the location of the speaker and the reference array are provided. The reference array is chosen to be the one that is situated in the same area. In this work, the results are reported for the single-array track [9] where only the data recorded by the reference array is used for recognition. The results in this paper are obtained on the development set because the transcriptions of the evaluation set are not publicly available at the time of this submission. Utterances having overlapped speech are not excluded from the training and the development sets.

5. DATA AUGMENTATION

Training data can be augmented to avoid over fitting and improve the robustness of the models [16]. Generally, adding more training data helps improving system's performance. In this work, we apply the speed perturbation based data augmentation technique [16] to increase the amount of training data of CHiME-5 corpus which consists of both binaural microphone data and Kinect microphone data (see section 4.3). The speed perturbation technique creates new data by resampling the original data. Two additional copies of the original training set are created by modifying the speed of speech to 90% and 110% of the original rate. The whole training set after data augmentation is 3 times larger than the original training set. The total amount of speech for training is around 501 hours. Due to the change in the length of the signals after resampling, the start and end times of the utterances are automatically updated by scaling the original start and end times with the resampling rates.

6. SPEECH RECOGNITION SYSTEM

6.1. Front-end processing

Acoustic features are extracted from the training and development data for training and testing of ASR systems. In the training set, individual speech signals from each microphone in each Kinect microphone array are used directly. In the development set using speech from the reference microphone array, speech signals from four microphones in the microphone array is processed with a weighted delay-and-sum beamformer (BeamformIt [23]) for enhancement prior to features extraction. In this paper, we do not use the development set consisting of speech recorded with the binaural microphones [9] as the focus is on the recognition of distant conversational speech.

FBANK and STE features of 40 dimensions are extracted from speech utterances which are obtained from the manual annotations by human on the binaural microphone data. In this work, the FBANK features are extracted in a conventional manner as follows: speech signal is first pre-emphasized by using a filter having a transfer function $H[z] = 1 - 0.97z^{-1}$. Speech frames of 25 ms are then extracted every 10 ms and multiplied with Hamming windows. DFT is used to transform speech frames into spectral domain. Sums of the element-wise multiplication between the magnitude spectrum and the Mel-scale filter-bank are computed. The FBANK coefficients are obtained by taking logarithm of these sums. The STE features are extracted using a filter-bank of $M = 40$ Gammatone filters (see section 3).

The features are extracted from speech utterances which are located in long audio sequences by using the provided start and end times. Utterance-level mean normalization is then applied. Both FBANK and STE features are augmented with 3-dimensional pitch features which include the value of pitch, delta-pitch and the probability of voicing at each frame [24, 9]. In this work, the FBANK and pitch features are extracted using the Kaldi speech recognition toolkit [25].

6.2. End-to-end ASR architecture

Hybrid CTC/attention end-to-end ASR systems [3] are built using the ESPnet toolkit [17]. The system architecture is depicted in Fig. 3. The architecture in this paper uses a shared encoder which consists of the initial layers of the VGG net architecture (deep convolutional neural network (CNN)) [26] followed by a 4-layer pyramid bidirectional long short-term memory (BLSTM with subsampling) [15], as in [27]. Here we use a 6-layer CNN architecture which consists of two consecutive 2D convolutional layers followed by one 2D Max-pooling layer, then another two 2D convolutional layers followed by one 2D max-pooling layer. The 2D filters used in the convolutional layers have the same size of 3×3. The max-pooling layers have patch of 3×3 and stride of 2×2. The 4-layer BLSTM has 320 cells in each layer and direction, and linear projection is followed by each BLSTM layer. The subsampling factor performed by the pyramid BLSTM is 4 [27].

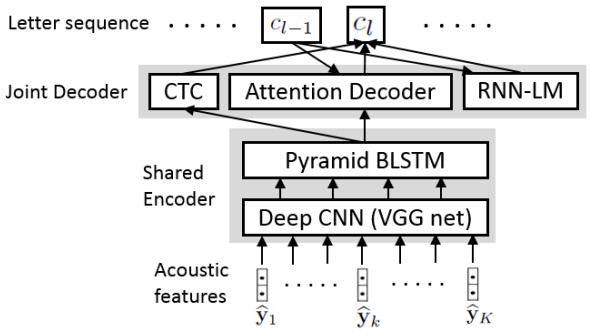

Fig. 3. Hybrid CTC/attention architecture [27, 3] of the end-to-end ASR systems used in this paper.

In this paper, location-based attention mechanism [2] is used in the hybrid CTC/attention architecture. This mechanism uses 10 centered convolution filters of width 100 to extract the convolutional features. The decoder network is a 1-layer LSTM with 300 cells. The hybrid CTC/attention architecture is trained within a multi-objective training framework by combining CTC and attention-based cross entropy to improve robustness and achieve fast convergence [17]. The training is performed with 15 epochs using the Chainer deep learning toolkit [28]. The CTC and attention-based cross entropy have equal weights (0.5) when being combined.

During joint decoding, CTC and attention-based scores are combined in a one-pass beam search algorithm [17]. A RNN language model (RNN-LM), which is a 1-layer LSTM, is trained on the transcriptions of the training data. This RNN-LM is used in the joint decoding where its log probability is combined with the CTC and attention-based scores [17]. The weight of the RNN-LM's log probability is set to 0.1 and the beam width is set to 20 during decoding.

6.3. System combination

As the training of hybrid CTC/attention end-to-end ASR systems does not use alignment information, these systems are not able to provide outputs with exact monotonic time-aligned information [3, 17]. System combination techniques, for instance lattice combination based on Bayes risk minimization [18] or ROVER (recognizer output voting error reduction) [19], use time-aligned information in the ASR outputs produced during the decoding to combine the ASR outputs. Without time-aligned information, it is not straightforward to apply these techniques to combine attention-based end-to-end ASR systems. In this paper, we propose a simple method to combine the hypotheses produced by ASR systems using FBANK and STE features. The method is described as follows.

Assume that we have two hypotheses produced by the end-to-end ASR systems using FBANK and STE features, respectively, for one speech utterance. From these two hypotheses, we will select the one which has higher output decoding score $\hat{s}$. The decoding score $s(C)$ is defined as the weighted sum of the CTC score, the attention-based score and the log probability of the RNN-LM [3, 17] given a letter sequence $C \in \mathcal{U}$ where $\mathcal{U}$ is a set of distinct letters. The output decoding score $\hat{s}$ is computed based on the letter sequence $\hat{C}$ which maximizes the decoding score over all possible letter sequences $C \in \mathcal{U}$. The letter sequence $\hat{C}$ is also the hypothesis output by the system. In short, $\hat{s} = s(\hat{C})$ where $\hat{C} = \arg\max_{C \in \mathcal{U}}\{s(C)\}$. The hypothesis selection method that we propose here is simple compared to other hypothesis selection methods, for instance [29]. This simple system combination method, which is denoted as "Combination" in the result tables (see section 7), is able to provide lower WER than those of the individual ASR systems.

To know how far we can reduce the WER if we are able to select the best hypothesis produced by the systems using FBANK and STE features for each utterance, we select the one which produces lower word error rate (WER) computed by using the manual transcription of the utterance. It should be noted that the hypothesis selection performed in this way is used only for reference because the transcription is not known in real condition. This method will be denoted as "Reference" in the result tables (see section 7).

7. EXPERIMENTAL RESULTS

Experimental results in terms of WERs are shown in Tables 1 and 2. The WERs for each session (party) in the development set and room conditions are also shown. Table 1 shows the results of the systems trained on the original training data set. In the present work, the baseline end-to-end system using FBANK features has a WER of 90.1% on the development set. The WER of the baseline end-to-end system introduced by the challenge organizers on the same development set was 94.7% [9]. Using the same setup with the system using FBANK features, the system using STE features provides 2.0% relative WER reduction compared to the system using FBANK features. Combing the hypotheses produced by the systems using FBANK and STE features using the proposed hypothesis selection method (see section 6.3) provides 4.3% relative WER reduction compared to the systems using FBANK features.

Table 1. Performance (WER, in %) of the ASR systems using FBANK and STE features. Both FBANK and STE features are augmented with pitch features. The systems are trained on the original training set without data augmentation.

Features	Session	Kitchen	Dining	Living	Overall
FBANK	S02	96.2	94.1	89.6	90.1
	S09	88.2	86.5	82.5	
STE	S02	96.1	89.1	87.0	88.3
	S09	89.4	84.7	81.6	
Combination	S02	94.0	88.0	85.8	86.2
	S09	83.8	82.0	77.9	
Reference	S02	92.3	86.6	82.9	84.1
	S09	82.1	80.1	76.0	

Table 2. Performance (WER, in %) of these ASR systems when the original training set are augmented by using speed perturbation based data augmentation technique.

Features	Session	Kitchen	Dining	Living	Overall
FBANK	S02	94.3	86.7	84.8	85.4
	S09	83.8	80.3	76.1	
STE	S02	92.8	85.0	81.6	84.2
	S09	82.9	82.0	77.6	
Combination	S02	91.2	83.0	79.9	81.4
	S09	78.6	78.2	72.2	
Reference	S02	88.9	80.4	77.1	79.0
	S09	77.1	75.3	70.3	

Tab. 2 shows the WERs of the systems trained on the augmented training set obtained with speed perturbation (see section 5). Data augmentation provides 5.2% and 4.6% relative WER reductions for the systems using FBANK and STE features, respectively, compared to the respective systems trained on the original training data. The system using STE features has 1.4% relative WER lower compared to the system using FBANK feature. Combining the systems using FBANK and STE features using the proposed hypothesis selection method provides 4.7% relative WER reductions compared to the system using FBANK features trained on the augmented training set.

It can be noticed from the results of the "Reference" method that the WER could be further reduced if the hypotheses produced by the systems using FBANK and STE features could be better combined. Improving the performance of the proposed hypothesis selection method to achieve that of the "Reference" method is one direction for future work.

8. CONCLUSION

This paper investigated the use of STE features and speed perturbation based data augmentation in end-to-end ASR of distant multi-microphone conversational speech in everyday home environments. A simple hypothesis selection method was proposed for combining the end-to-end systems using FBANK and STE features. The investigated techniques helped improving end-to-end ASR performance on a challenging corpus which was used for the CHiME 2018 speech separation and recognition challenge [9]. STE features are extracted with a different method and are complementary to FBANK features. This complement is the source of up to 4.7% relative WER reduction obtained when combining the ASR systems using these two features. The accumulated relative WER reduction obtained by both data augmentation and combining systems using the two features is 9.7%.

9. REFERENCES

[1] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. of the 31st International Conference on Machine Learning*, Beijing, China, June 2014, pp. 1764–1772.

[2] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 577–585.

[3] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 1240–1253, December 2017.

[4] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[5] S. Thomas, S. Ganapathy, and H. Hermansky, "Phoneme recognition using spectral envelope and modulation frequency features," in *Proc. IEEE ICASSP*, Taiwan, April 2009, pp. 4453–4459.

[6] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 4117–4120.

[7] C.-T. Do and Y. Stylianou, "Improved automatic speech recognition using subband temporal envelope features and time-delay neural network denoising autoencoder," in *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 3832–3836.

[8] S. Ganapathy and M. Harish, "Far-field speech recognition using multivariate autoregressive models," in *Proc. INTERSPEECH*, Hyderabad, India, September 2018, pp. 3023–3027.

[9] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proc. INTERSPEECH*, Hyderabad, India, September 2018, pp. 1561–1565.

[10] D. Palaz, R. Collobert, and M.-M. Doss, "End-to-end phoneme sequence recognition using convolutional neural networks," in *Proc. NIPS Deep Learning Workshop*, 2013, pp. 1–8.

[11] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," in *arXiv preprint arXiv: 1609.03193*, 2016.

[12] A. Tjandra, S. Sakti, and S. Nakamura, "Attention-based wav2text with feature transfer learning," in *Proc. 2017 IEEE Automatic Speech Recognition and Understanding Workshop*, Okinawa, Japan, December 2017, pp. 309–315.

[13] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, "End-to-end speech recognition from the raw waveform," in *Proc. INTERSPEECH*, Hyderabad, India, September 2018, pp. 781–785.

[14] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, B. Casper, J. Catanzaro, Q. Cheng, G. Chen, et al., "Deep speech 2: end-to-end speech recognition in english and mandarin," in *Proc. of the 33rd International Conference on Machine Learning*, New York, USA, June 2016, pp. 173–182.

[15] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *Proc. IEEE ICASSP*, Shanghai, China, March 2016, pp. 4960–4964.

[16] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 3586–3589.

[17] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: end-to-end speech processing toolkit," in *Proc. INTERSPEECH*, Hyderabad, India, September 2018, pp. 2207–2211.

[18] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, vol. 25, no. 4, pp. 802–828, 2011.

[19] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Dec 1997, pp. 347–354.

[20] D. B. Paul and J. M. Barker, "The design for the Wall Street Journal-based CSR corpus," in *HLT '91 Proceedings of the workshop on Speech and Natural Language*, New York, USA, February 1992, pp. 357–362.

[21] M. Slaney, *Auditory Toolbox (version 2)*, Interval Research Corporation Technical Report #1998-010, 1998.

[22] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 207, no. 5234, pp. 303–304, 1995.

[23] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2011–2023, 2007.

[24] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014, pp. 2513–2517.

[25] 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 *Proc. IEEE ASRU 2011*, Hawaii, USA, December 2011.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. International Conference on Learning Representations*, 2015.

[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 *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 949–953.

[28] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proc. of NIPS Workshop on Machine Learning Systems (LearningSys)*, 2015.

[29] V. Soto, O. Siohan, M. Elfeky, and P. Moreno, "Selection and combination of hypotheses for dialectal speech recognition," in *Proc. IEEE ICASSP*, Shanghai, China, March 2016, pp. 5845–5849.