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

◇ Investigate the use of subband temporal envelope (STE) features and speed perturbation based data augmentation in end-to-end speech recognition.

◇ Experiments are performed on CHiME-5 corpus of distant conversational speech in everyday home environments.

◇ STE features yields better performance than the conventional log-Mel filter-bank (FBANK) features.

◇ Data augmentation is used with both features and yields up to 5.2% relative improvement.

◇ Combining systems using FBANK and STE features yields additional 4.7% relative improvement.

■ Subband temporal envelope (STE) features

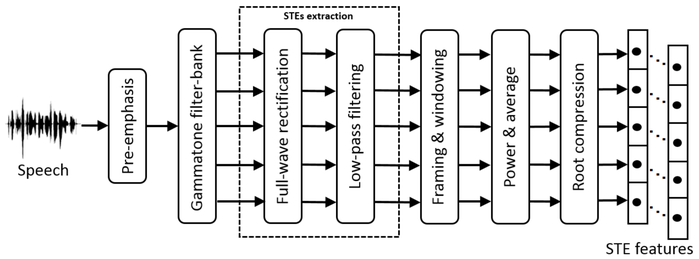

Figure 1: Algorithm for extracting the STE features.

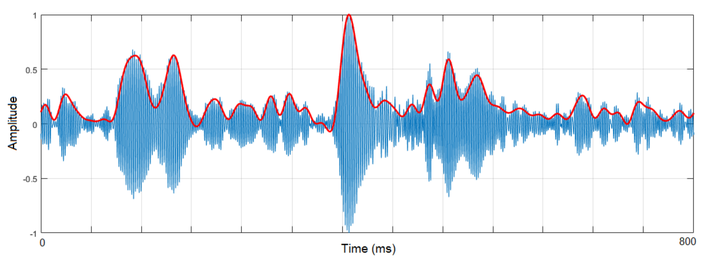

Figure 2: Slowly-varying STE (red curve) extracted from the 10th subband signal of a speech segment of 800 ms.

◇ STE features are computed using 40 Gammatone filter-banks.

◇ 40-dimensional FBANK features are extracted using Kaldi.

◇ Both features are augmented with 3-dimensional pitch features extracted with Kaldi.

■ CHiME-5 speech corpus

☐ 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.

◇ Natural conversational speech from a dinner party of 4 participants was recorded.

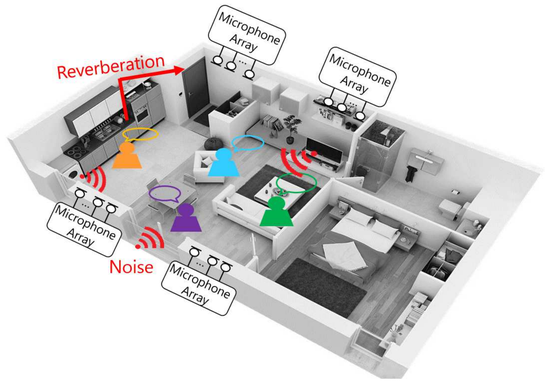

Figure 3: An illustration of the CHiME-5 corpus recording scenario.

☐ Data for training and test

◇ The data used for training combines both left and right channels of the binaural microphone data and a subset of all Kinect microphone data from 16 parties (167 hours of speech).

◇ Each of the development and evaluation sets is created from 2 parties of around 4.5 and 5.2 hours of speech, respectively.

■ Data augmentation

◇ Speed perturbation based data augmentation is used to create new data by resampling the original data.

◇ Two additional copies of the original training sets are created by modifying the speed of speech to 90% and 110% of the original rate.

■ Speech recognition experiments

☐ ASR systems

◇ Weighted delay-and-sum beamformer (BeamformIt) is applied on the test set prior to features extraction.

◇ Hybrid CTC/attention end-to-end architecture is used.

☐ System combination

◇ A simple hypothesis selection method is proposed to combine systems using FBANK and STE features. The selection is based on the weighted sum of the CTC, attention, and RNN-LM scores.

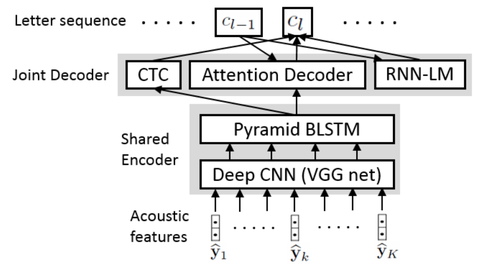

Figure 4: Hybrid CTC/attention end-to-end architecture [ESPnet].

☐ Experimental Results

Table 1: Performance (WER, in %) of the ASR systems trained on the original training set.

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	

■ Conclusion

◇ Using STE features and speed perturbation based data augmentation in end-to-end ASR of distant conversational speech was effective.

◇ Experiments were performed on a challenging corpus used for the CHiME 2018 speech separation and recognition challenge.

◇ The accumulated relative WER reduction obtained by data augmentation and combining systems using the two features was 9.7%.