Audio Codec Simulation based Data Augmentation for Telephony Speech Recognition

Thi-Ly Vu*, Zhiping Zeng*, Haihua Xu*, and Eng-Siong Chng*
* Nanyang Technological University, Singapore
E-mail: {tlvu, zengzp, haihuaxu, ASESChng}@ntu.edu.sg

***Abstract*—Real telephony speech recognition task is challenging due to 1) diversified channel distortions and 2) limited access to the real data because of the data privacy consideration. In this paper, assuming no real telephony data are available, we employ diversified audio codecs simulation based data augmentation method to train telephony speech recognition system. Specifically, we assume only wide-band 16 kHz data are available, and we first down-sample the 16 kHz data to the 8 kHz data; we then pass the down-sampled data through various categories of audio codecs to simulate the real channel distortion. As a result, we train our speech recognition with such distorted data. To analyze the effectiveness of different audio codec simulation methods, we classify them into three main categories according to their distortion severity, in terms of their spectrogram analysis. We conduct experiments on various real telephony test sets to show the effectiveness of the proposed data augmentation method. The result shows that the real data is more close with highly distorted simulation data, since the model with highly distorted data reduce the Word-Error-Rate 7.28% - 12.78% compared to the baseline.**

I. Introduction

Spontaneous speech recognition accuracy has been remarkably improved in recent years thanks to the advent of Deep Neural Network (DNN) modeling techniques, and the usage of big training data [1] [2]. However, telephony speech recognition is still challenging. First of all, it is hard to obtain sufficient domain specific real telephony data to train acoustic models due to data privacy consideration. Secondly, telephony speech itself is usually highly distorted due to diversified channel codecs applied. Furthermore, if the data itself is contaminated with other environmental additive noise or reverberant noise, recognition results over a domain specific telephony data will be significantly degraded [3], [4], [5], [6], [7].

In this paper, assuming no real telephony Singapore English data available, we employ various kinds of audio codecs simulation based data augmentation method as in [4], [8], [9], [10], [11], training telephony speech recognition systems to recognize real telephony test data. Specifically, to obtain 8 kHz telephony training data, we down-sample 16 kHz data to 8 kHz data, we then pass the down-sampled data through various categories of audio codecs to simulate real channel distortion.

To maximize the effectiveness of our codecs simulation based data augmentation method, we collect 27 codecs in total, analyze their spectrograms, and then classify the codec into three categories according to the severity of the distortion displayed from sepctrograms and Mean Opinion Score (MOS). MOS is a subjective measure of sound quality from 1 to 5. After that, we train a speech recognition system for each codecs category, as well as one using the simulated data with the mixed codecs.

Our work is motivated by a real project requirement, where little real telephony training data is available due to the data secret policy requirement. However, we manage to get several real telephony test data sets for evaluation. The main contribution of the paper lies in the following aspects: 1) We collect extensive audio codecs, which is up to 27 in total. To the best of our knowledge, it is a comprehensive codec set compared with the previous works in [3], [4], [8], [5], [10]. 2) We use the completely mismatched wide-band data to train 8 kHz telephony speech recognition systems, of which the effectiveness of the proposed method is evaluated on the real telephony speech data. 3) We use up to four real telephony test data sets obtained from different sources. This ensures the effectiveness of the proposed method.

The paper is organized as follows. Section II introduces codec list we collected and shows our spectrogram analysis. Section III describes the details of our work doing audio codec simulation. Section IV is the description of the training and evaluation data we employed. Section V presents the experimental setups and the results. Finally, Section VI concludes the work.

A. Related works

GSM, G711, G723.1 and MPEG coders are investigated [4], [5]. The researches show that the GSM full rate and MPEG (below 32kbit) degrade the speech recognition performance significantly whereas G711 and G723.1 do not have such effect. They suggested that we could keep the acoustic model trained on clean speech and learn a linear transformation f between "clean" and "degraded" signals. During the recognition stage, the inverse transformation f^{-1} would be applied to the degraded test signal to reduce mismatch between training and test. This study also pointed out that the packet loss during transmission, the biggest source of degradation, was not in their scope. It was presented in [6], [8]. The influences caused by packet loss were reduced by strategies to recover lost information, or the relative weight of the language and acoustic model is changed according to the packet loss rate.

The research in [10] proposed a scheme adding both environmental noises and codecs condition to train the speech system, and evaluated with speaker recognition. The codecs was grouped by the codec type, and adding as the additional to the noisy data by environment.

Jiří Málek studied the effects of single codecs and mixed of several codecs on the speech system performance in [11]. Their study including experiments with single codec, and mixed the different number of codecs. The mixed codecs help mitigate the deteriorated performance due to training on off-domain data.

Inspired by above researches in speech and audio codecs, we examine the comprehensive list of the codecs using FFMPEG to simulate these codecs. The list including codecs used in landline telephone communication, satellite/radio transmission, and Voice over IP. Using such simulated data, we train the code-switch speech recognition systems [12], [13] using SEAME corpus [14]. Different from almost researches about codecs, we categorize the list of codecs based on the MOS, and the level of distortion in the spectrogram. We also evaluate the telephony speech recognition with several real telephony data sets. These evaluation sets are diversified in domains, number of speakers, and the quality of the audio. These real evaluation sets give us more detail and valuable analysis of the model performance. Our data augmentation work-flow is illustrated in Figure 1.

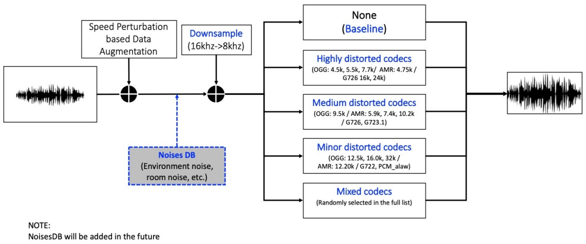

Speed Perturbation based Data Augmentation
Downsample (16khz->8khz)
Noises DB (Environment noise, room noise, etc.)
None (Baseline)
Highly distorted codecs (OGG: 4.5k, 5.5k, 7.7k/ AMR: 4.75k / G726 16k, 24k)
Medium distorted codecs (OGG: 9.5k / AMR: 5.9k, 7.4k, 10.2k / G726, G723.1)
Minor distorted codecs (OGG: 12.5k, 16.0k, 32k / AMR: 12.20k / G722, PCM_alaw)
Mixed codecs (Randomly selected in the full list)
NOTE:
NoisesDB will be added in the future

Fig. 1: Data augmentation scheme in our study

II. Codecs and analysis

A. Codecs

The comprehensive list of 27 speech and audio codecs (see in the Table I), are studied in this paper.

- **Landline** includes mu/A-law companding. It also includes Adaptive-Differential PCM (ADPCM) coding following the ITU G.726 standard, allowing for 16, 32, 48 and 64 kbps rates.
- **Cellular** includes two major cellular telephony codecs, namely the Global System for Mobile Communications (GSM) and narrow-band and wide-band Advance Multi-rate (AMR-NB and AMR-WB) codecs. The GSM standard supports four different but similar compression technologies to analyse and compress speech. These include full-rate, enhanced full-rate (EFR), adaptive multi-rate (AMR), and half-rate.
  - The full-rate allowing for 13 kbps rate uses linear prediction coding with regular pulse excitation.
  - EFR (Enhanced Full Rate) uses ACELP (Algebraic Code Excited Linear Prediction)
  - HR (Half Rate) uses CELP-VSELP (Code Excited Linear Prediction – Vector Sum Excited Linear Prediction)
  - AMR-NB is a multi-rate speech codec using Algebraic Code-Excited Linear Prediction (ACELP) at 4.75-12.2 kbps. AMR-WB, following the ITU G.722.2 specification, is the wide-band variant of AMR, coding speech signals up to 7 kHz using bit-rates from 6.6 to 23.8 kbps.
- **Satellite/Radio** includes three codecs (ITU G.728, Continuously Variable Slope Delta (CVSD), and Codec2) that are used in satellite and radio telecommunication systems.
- **VoIP** includes the ITU G.729, ITU, G726, G723.1, G722 and G711 standards besides SILK and SILK-WB, former Skype now open-source codecs.
  - ITU G.729a is a narrow-band low-complexity codec based on the Code-Excited variant of ACELP (CS-ACELP), operating at 8 kbps.
  - G726 is an improved version of G.721 and G.723 (different from G.723.1).
  - G.723.1 includes two variants. The bitrate of the first variant is 6.4 kbit/s and the MOS is 3.9. The bitrate of the second variant is 5.3 kbit/s with MOS=3.7.
  - The ITU G.722 is a wide-band audio codec based on sub-band ADPCM allowing 48, 56 and 64 kbps rates. It is used for voice over IP and radio broadcasters.
  - G711 is Pulse code modulation (PCM) of voice frequencies, that was introduced by ITU in 1972 for use in digital telephony. The codec has two variants: A-Law is being used in Europe and in international telephone links, u-Law is used in the U.S.A. and Japan. G.711 uses a logarithmic compression. It squeezes each 16-bit sample to 8 bits. The MOS value is 4.2.
  - The Opus (OGG) format is based on a combination of the full-bandwidth CELT format and the speech-oriented SILK format, both heavily modified: CELT is based on the MDCT that most music codecs use, using CELP techniques in the frequency domain

for better prediction, while SILK uses linear predictive coding (LPC) and an optional Long-Term Prediction filter to model speech.

TABLE I: List of codecs used in our study

	Codec information		
	Type	Bitrates (kBits/sec)	SampleRates (kHz)
G726	ADPCM	16/24/32/40	8
GSM 06.10 (Full-rate)	GSM	6.70, 7.40, 7.95, 10.20, 12.20	8
GSM 06.20 (Half-rate)	GSM	4.75, 5.15, 5.90, 6.6	8
AMR-NB	ADPCM	4.75, 5.90, 7.40, 10.20, 12.20	8
Opus / SILK (VOIP)	OGG	4.5, 5.5, 7.7, 9.5, 12.5, 16.0, 32.0	8
G722	ADPCM	64	16
G723.1	ADPCM	6.3	8

B. Codecs analysis

In this study, we studied the group of codecs based on the level of distortion to the spectrogram and the MOS. We grouped into 4 categories: highly distorted codecs, medium distorted codecs, minor distorted codecs, and finally, the mixed set (which codec in the comprehensive list will be randomly chosen and applied to the clean audio file).

The example spectrogram of each group was illustrated in Figure 2. As you can observed in this Figure, the harmonic and high frequencies in the (b) is much more distorted than the spectrogram in the (c) and (d). There are also more noises in higher frequencies.

The level of distortion in the spectrogram also reflected in the quality of the audio sound. In the highly distorted codecs group, the sound is very bad, creaky like frame drop or clipping. In the real data, the bad quality of sound may cause by the codec or the bad transmission communication. In the medium distorted codecs group, these noises are less and the sound is better. And the sound in the minor distorted codecs group is not much different with the clean audio.

III. Audio codec simulation for speech recognition

We use FFMPEG to generate the simulated data. Since the binary installation of the FFMPEG may not support all the libraries, we recommend you compile FFMPEG from source with `libfdk-aac`, `libopus`, `libmp3lame`, `libx264` and `amr` libraries enabled.

To build the Speech Recognition with simulated data, we convert clean training data through the codec simulation, as demonstrated below:

```
ffmpeg -i input.wav -c:a libopus \
    -b:a 5.5k -ar 8000 output_550.ogg
ffmpeg -i output_550.ogg \
    -ar 8000 output_550_ogg.wav
```

In practice, we generate simulated data in accordance with kaldi-format, so we do not generate the intermediate file, use pipe instead.

To train the augmented model with highly distorted data, we compose the list of highly distorted codecs, and select randomly from this list, to apply to clean audio, to generate high distorted training data. Similarly, we have medium distorted training data, minor distorted data, mix distorted training data. In the mix distorted training data, we select codec randomly from the full list.

IV. Data

A. Training data

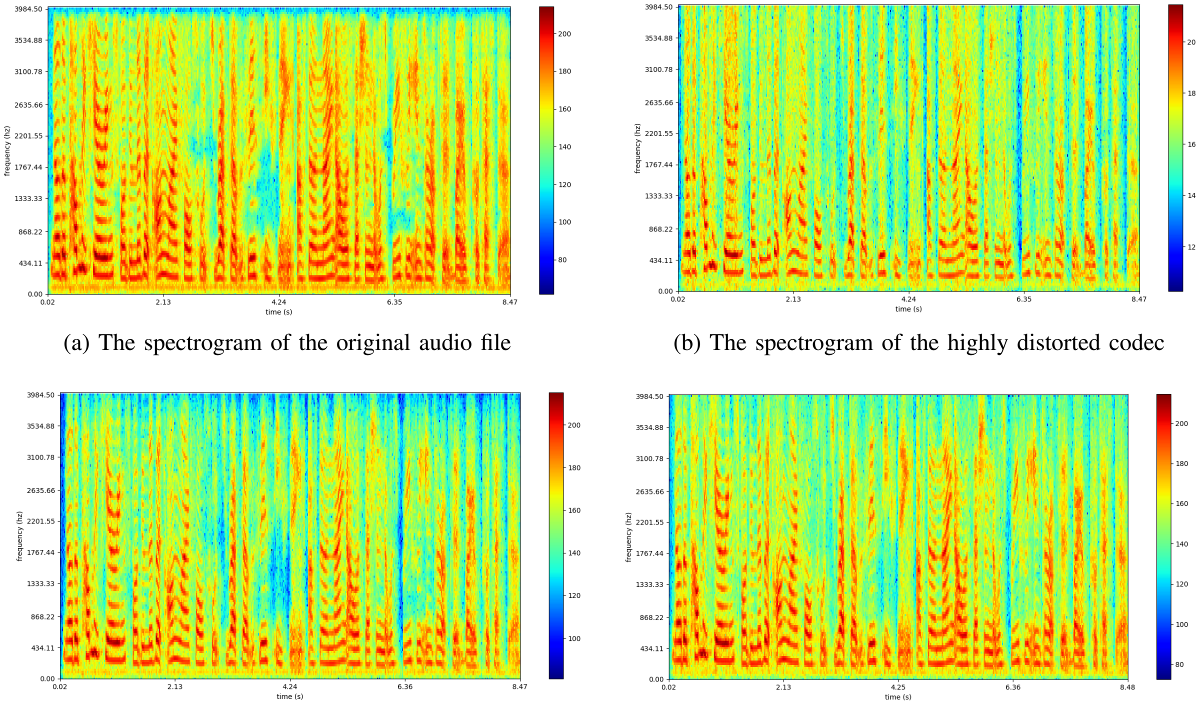
We use about 100 hours of SEAME corpus to train our speech recognizer. The SEAME corpus is a microphone based spontaneous conversational bilingual speech corpus, of which most of the utterances contain both English and Mandarin in Malaysia and Singapore areas [14]. From the distribution of speakers, we can see the SEAME corpus is generally biased with Mandarin. Furthermore, they find Singaporean speakers normally have more English words in their utterances, while Malaysian speakers are more likely to converse with utterances dominated by Mandarin.

B. Test data

To evaluate the proposed methods, we define eight evaluation data sets. Two of them are extracted in the same domain with training set, each are randomly selected from about 10 gender balanced speakers. However they are defined differently, and one is dominated by Mandarin, named as Dev_{man}, and the other is dominated by English, as Dev_{sge}. These two proposed "biased" data sets [13], [15] would give more clues to show the effectiveness of each proposed method on each individual languages. We also generate simulated data from these two dev sets with high distorted codecs, namely $\text{Dev}_{\text{man-noisy}}$, and $\text{Dev}_{\text{man-noisy}}$ respectively.

Four evaluation sets are real telephony conversation. These evaluation set are out of domain with the training set. All these real telephony evaluation sets are dominated by English. The speakers in these sets are diversified, from 16 to more than 400 speakers, uttered by speakers mainly from Singapore and Malaysia. The FreeTalk-2016, CallCenter, and DailyConversation test sets are the real telephony data collected in the call center and telecommunication companies/organizations, while the FreeTalk-2019 is the test set recorded by people in the data company.

More statistics about these evaluation sets are described in Table II.



(a) The spectrogram of the original audio file

(b) The spectrogram of the highly distorted codec

(c) The spectrogram of the medium distorted codec

(d) The spectrogram of the minor distorted codec

Fig. 2: Illustrate the spectrogram of audio files with different level of distortion

TABLE II: The statistics of data sets

Data Type	Data Set	#SPKS	Length (hrs)	Remark (8kHz)
Train	SEAME$_{\text{train}}$	134	101.1	Clean
Dev	Dev$_{\text{sge}}$	10	4	Clean
	Dev$_{\text{man}}$	10	7.5	Clean
	Dev$_{\text{sge-noisy}}$	10	4	Noisy simulated
	Dev$_{\text{man-noisy}}$	10	7.5	Noisy simulated
Eval	FreeTalk-2016	20	3.73	Real tel scene, Noisy
	CallCenter	406	3.06	Real tel scene, Noisy
	FreeTalk-2019	30	2.91	Tel scene, Noisy
	Daily-Conversation	16	0.38	Real tel scene, Noisy

V. Experiment Setup

We report results on speech recognition task in English and Mandarin. Baseline model is trained on the 100 hour SEAME data set [15]. All the experiments using 13-layers with 1024 hidden units TDNN-F [16] run in Kaldi toolkit [17] . All data are sampled at 8kHz. Lexicon contain 39k English words and Chinese characters. The tri-gram language model only use the train data transcription to make graph in decoding. We also train all ASR systems with standard 3-way speed perturbation data augmentation [18] using factors of 0.9, 1.0 and 1.1.

A. *Baseline models*

The baseline model was trained with down-sampled SEAME data (8kHz).

B. *Augmented models*

The augmented models were trained with augmented SEAME data. In the first set of experiments, we apply the single condition to the whole training data set. For example, in the highly-distorted model, we select randomly codec from the highly-distorted codecs, and apply to the clean data. In the mixed model, we select codec from the full list of codecs. As the results, we have 4 models in the first set of experiments, namely Highly Distorted Codecs (HighDC), Medium Distorted Codec (MediumDC), Minor Distorted Codec (MinorDC) and finally, Mixed Codecs (MixCodec). In the second set of experiments, we pass the clean data through another data augmentation process: speed perturbation based process, so we have HighDC-SP (Highly Distorted Codecs and Speed Perturbation), MediumDC-SP (Medium Distorted Codecs and Speed Perturbation), MinorDC-SP (Minor Distorted Codecs and Speed Perturbation), and MixCodec-SP (Mixed Distorted Codecs and Speed Perturbation) models.

In the end, we have 9 models (BASELINE model and 8 augmented models).

C. *Results*

We reported the Word Error Rate (WER) with eight evaluation sets, and results are summarized in 2 tables.

Table III contains WER of the codecs augmented model with real telephony data. The WER of the highly distorted codec model (HighDC) is the lowest, drop significantly, from **7.28%** to **12.78%** compared to baseline model. This indicates that the highly distorted simulation data is much more close to the real telephony data. The WER of other augmented models reduces consistently from highly to minor distorted codecs group. The mixed codecs model is very close to the HighDC, and is robust to all test sets in our study. The second rows in the Table show the results with augmented model, adding speed perturbation based augmentation method (HighDC-SP, MediumDC-SP, MinorDC-SP, MixCodec-SP). The speed perturbation data augmentation method does a little help but inconsistent with the real data (the improvements are highlight in the bold text).

TABLE III: WER of models with codecs and real telephony data

Models	**FreeTalk-2016**	**CallCenter**	**Freetalk-2019**	**Daily-Conversation**
BASELINE	64.78	62.59	65.61	69.02
HighDC	**55.09**	53.30	52.83	**61.74**
HighDC-SP	56.54	**53.07**	**51.73**	63.27
MediumDC	59.78	56.62	61.27	64.82
MediumDC-SP	**56.72**	**53.75**	**55.31**	**63.98**
MinorDC	**59.48**	**57.15**	59.38	**65.37**
MinorDC-SP	60.17	57.15	**58.73**	66.69
MixCodec	**55.64**	53.63	53.64	**62.52**
MixCodec-SP	56.55	**52.33**	**53.11**	63.57

Table IV contains WER of the codecs augmented model with dev data sets. The BASELINE model is the best for dev sets. However, the HighDC are the best for codec simulated dev sets. The results with simulated dev sets are consistently with the real telephony, HighDC model improves for $\text{Dev}_{\text{sge-noisy}}$ **8.02%**, from 40.50% to 32.48% , for $\text{Dev}_{\text{man-noisy}}$ drop **8.4%**, from 33.27% to 24.87% . This indicates that codec based data augmentation method helps to improve the speech recognition performance.

TABLE IV: WER of models with codecs and dev data

Models	**Dev_{sge}**	**$\text{Dev}_{\text{sge-noisy}}$**	**Dev_{man}**	**$\text{Dev}_{\text{man-noisy}}$**
BASELINE	26.79	40.50	19.89	33.27
HighDC	28.51	32.48	21.46	24.87
HighDC-SP	**27.75**	**31.63**	**20.92**	**24.00**
MediumDC	27.66	36.17	20.64	28.90
MediumDC-SP	**27.03**	**35.13**	**20.18**	**28.13**
MinorDC	27.16	37.81	20.51	30.53
MinorDC-SP	**26.71**	**37.33**	**19.9**	**29.99**
MixCodec	27.56	33.10	20.58	25.68
MixCodec-SP	**27.11**	**32.2**	**20.05**	**24.97**

VI. Conclusion and Future Work

In this paper, we proposed the data augmented scheme that use codecs and speed perturbation based method to build the telephony speech recognition system. The model trained with proposed data augmented scheme using highly distorted codecs gets better result with the real telephony data, suggest codecs in this group is more closer to the real world setting. The model trained in a multi-condition (mixed codecs) fashion yields comparable performance to specialized model trained for highly distorted codecs group and is robust to unseen test conditions. In the future, we would like to explore the data augmented scheme with more environment noises, reverberation [19] to handle various conditions in the real world as in [20], [21].

VII. Acknowledgment

Chng Eng Siong is supported by the Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University, Singapore.

Xu Haihua is supported by TL@NTU, Nanyang Technological University, Singapore.

References

[1] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[2] Y. Zhou, C. Xiong, and R. Socher, "Improved regularization techniques for end-to-end speech recognition," *arXiv preprint arXiv:1712.07108*, 2017.

[3] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 4. IEEE, 1996, pp. 2344–2347.

[4] J. M. Huerta and R. M. Stern, "Speech recognition from gsm codec parameters," in *Fifth International Conference on Spoken Language Processing*, 1998.

[5] L. Besacier, C. Bergamini, D. Vaufreydaz, and E. Castelli, "The effect of speech and audio compression on speech recognition performance," in *2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564)*. IEEE, 2001, pp. 301–306.

[6] P. Mayorga, L. Besacier, R. Lamy, and J.-F. Serignat, "Audio packet loss over ip and speech recognition," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 607–612.

[7] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[8] B. Milner and S. Semnani, "Robust speech recognition over ip networks," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 3. IEEE, 2000, pp. 1791–1794.

[9] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 9, pp. 1469–1477, 2015.

[10] M. Ferras, S. Madikeri, P. Motlicek, S. Dey, and H. Bourlard, "A large-scale open-source acoustic simulator for speaker recognition," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 527–531, 2016.

[11] J. Málek, J. Žd́anskỳ, and P. Červa, "Robust recognition of conversational telephone speech via multi-condition training and data augmentation," in *International Conference on Text, Speech, and Dialogue*. Springer, 2018, pp. 324–333.

[12] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A first speech recognition system for mandarin-english code-switch conversational speech," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4889–4892.

[13] P. Guo, H. Xu, L. Xie, and E. S. Chng, "Study of semi-supervised approaches to improving english-mandarin code-switching speech recognition," *arXiv preprint arXiv:1806.06200*, 2018.

[14] D.-C. Lyu, T.-P. Tan, E. S. Chng, and H. Li, "Seame: a mandarin-english code-switching speech corpus in south-east asia," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[15] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the end-to-end solution to mandarin-english code-switching speech recognition," *arXiv preprint arXiv:1811.00241*, 2018.

[16] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of INTERSPEECH*, 2018.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[18] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[20] A. Narayanan, A. Misra, K. C. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 441–447.

[21] I. Szoke, M. Skacel, L. Mosner, J. Paliesek, and J. Cernocky, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, 2019.