INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Robust Speech Recognition Via Anchor Word Representations

Brian King[1], *I-Fan Chen*[1], *Yonatan Vaizman*[2], *Yuzong Liu*[1], *Roland Maas*[1], *Sree Hari Krishnan Parthasarathi*[1], *Björn Hoffmeister*[1]

[1]Amazon, U.S.A.
[2]University of California, San Diego, U.S.A

{bbking,ifanchen,liuyuzon,rmaas,sparta,bjornh}@amazon.com, yonatanv@gmail.com

Abstract

A challenge for speech recognition for voice-controlled household devices, like the Amazon Echo or Google Home, is robustness against interfering background speech. Formulated as a far-field speech recognition problem, another person or media device in proximity can produce background speech that can interfere with the device-directed speech. We expand on our previous work on device-directed speech detection in the far-field speech setting and introduce two approaches for robust acoustic modeling. Both methods are based on the idea of using an anchor word taken from the device directed speech. Our first method employs a simple yet effective normalization of the acoustic features by subtracting the mean derived over the anchor word. The second method utilizes an encoder network projecting the anchor word onto a fixed-size embedding, which serves as an additional input to the acoustic model. The encoder network and acoustic model are jointly trained. Results on an in-house dataset reveal that, in the presence of background speech, the proposed approaches can achieve up to 35% relative word error rate reduction.

Index Terms: robust speech recognition, speaker adaptation, encoder-decoder network

1. Introduction

While automatic speech recognition (ASR) recently reached another milestone by achieving human parity on a telephony task [1], far-field ASR has proven to be more challenging and requires further innovation. To encourage innovation in this field, there have been several ASR challenges in the last few years. The REVERB [2], CHiME [3, 4, 5], and ASPiRE [6] tasks focus on the challenges with difficult acoustic conditions and mismatched data, while the MGB [7] task illustrates the difficulty in working with noisy transcripts.

In this paper, we study the problem of robustness in far-field speech recognition, specifically robustness to interfering background speakers and media speech, when a foreground speaker addresses a voice-controlled device uttering an "anchor word" at the beginning of an interaction. We will call the foreground speech from the desired talker "desired speech" . Consider the following interaction:

[speaker 1:] Computer, what is the weather in New York?
[speaker 2:] Take the dog out for a walk.

Here we consider "computer" as the anchor word, the utterance by speaker 1 as desired speech, and the utterance by speaker 2 as interfering speech. The aim of this paper is to use the salient information in the anchor word to better recognize the desired speech and ignore the other speech.

Much work has been done on the topic of noise-robust ASR; even within the confines of single channel audio, a number of feature and model space approaches have been proposed. For feature space, blind deconvolution through long-term averaging [8] (alternatively called the cepstral mean subtraction in the speech community [9]) is a standard "normalization" method to achieve robustness to channel noise. There are also methods that normalize in the spectral domain [10]. These methods assume that the noise and clean speech are uncorrelated and additive in the time domain. Normalization using the second order statistics, or the cepstral variance normalization, was proposed in [11]. Other methods, such as RASTA [12], exploit the human auditory perception characteristics to filter the feature trajectory.

In the context of the GMM-HMM acoustic model (AM), model space methods for robustness use the vector Taylor series (VTS) family of approximations [13] or maximum likelihood linear regression (MLLR) family of linear transforms [14, 15]. Note that several connections exist between the model space methods and the feature space methods [16]. With DNNs for AM, noise modeling can be less explicit. For example, an estimated auxiliary noise vector is employed with standard features in [17]. Furthermore, training methods, such as dropout, or activations, such as maxout, have been shown to be useful to provide robustness [18]. Novel DNN architectures such as network-in-network can also be useful for noise robustness, as seen in the NTT submission for CHiME3 [19].

Moving beyond the general issue of noise robustness to the specific issue of robustness to interfering speech, a body of work addressed them as crosstalk or as overlapped speech in meeting room ASR tasks [20]. Typically, these exploit multichannel audio to derive feature and model space segmentation methods [21, 22]. Segmented audio was then used for recognition [23]. Exploiting the flexibility of DNNs and drawing motivation from meeting room approaches for a robust segmentation, our earlier work on speech detection in the presence of background speech [24] explored two approaches, a model space method and a feature space method, to learn representations from the anchor word; both methods performed comparably for the speech detection task.

In this paper, we extend the two approaches presented for desired speech detection in [24] to acoustic modeling for desired speaker speech recognition. Our first method, termed anchored mean subtraction (AMS), extracts the anchor word mean to normalize the speech features in the utterance. The second method is based on an encoder-decoder model where the encoder network is trained jointly with an decoder AM to learn a representation from the anchor. Experimental results show both methods provide significant reductions in WER for utterances with background speech, while the model space approach consistently outperform the feature space approach.

The remainder of this paper is structured as follows: the AMS and encoder-decoder methods are reviewed in Section 2; the experimental setup and results are presented in Sections 3

http://dx.doi.org/10.21437/Interspeech.2017-1570

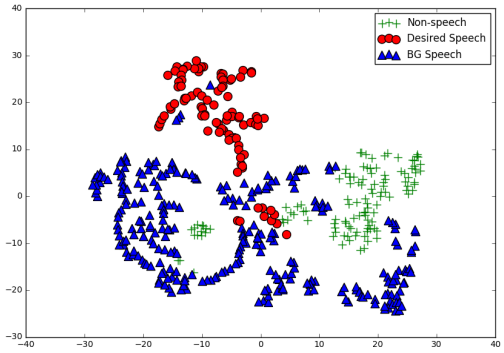

Figure 1: *Distribution of CMS-normalized features of an utterance with background speech. The distribution is visualized by mapping 64-dimensional log filter bank energy features to 2 dimensions using the t-SNE algorithm. It is clear that CMS-normalized features of the non-speech and background (BG) speech overlap significantly in the feature space.*

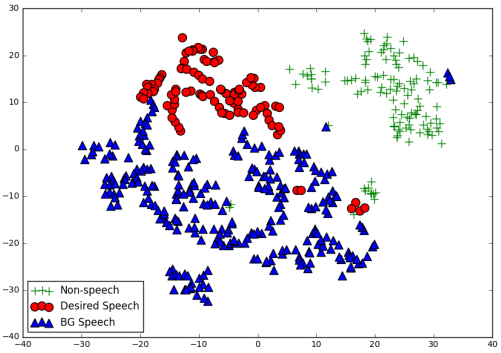

Figure 2: *Distribution of AMS-normalized features of the same utterance in Figure 1 using the t-SNE algorithm. Compared to Figure 1, AMS-normalized desired speech, background (BG) speech, and non-speech overlap less and are more separable in the feature space.*

and 4; and conclusions are drawn in Section 5.

2. Anchor Word Embedding

We briefly review two methods of embedding anchor word information into a DNN-based AM, which were originally proposed for speech detection [24] to distinguish desired speaker speech from background speech and noise.

2.1. Anchored Mean Subtraction

Mean normalization can improve the robustness of ASR systems [9, 10]. The channel transfer function on log frequency domain, H, can be computed as an online, recursive update [25], using the following:

$$\hat{H}_{k,n+1} = \alpha \hat{H}_{k,n} + (1-\alpha) X_{k,n}, \text{ for } 0 < \alpha \leq 1, \quad (1)$$

where $\hat{H}_{k,n}$ and $X_{k,n}$ denote the estimate of H and the log STFT magnitude of the observed far-field signal for frame n and dimension k, respectively. The parameter α is a hyper-parameter, and this estimator can be interpreted as a low-pass filter. Referring to this as causal mean subtraction (CMS), normalization of the observed signal can then be achieved by $\hat{S}_{k,n} = X_{k,n} - H_{k,n}$, where $\hat{S}_{k,n}$ denotes the log STFT magnitude of the estimated speech signal. CMS can cause desired and interfering speech features to become more similar, which is in opposition to our goal of recognizing only desired speech. For example, volume differences may be equalized [24]. Furthermore, with CMS, since the normalization factor applied to the current frame is affected by the previous frames, it can negatively impact the features: two identical signals will differ if their preceding frames differ. To address these issues, we utilize anchored mean subtraction (AMS) [24]. Unlike CMS, where the mean estimate $\hat{H}_n$ is updated every frame, in AMS, the mean $\hat{H}_{1:l}$ is calculated over the anchor word segment (spanning frames 1 to l), and is kept constant over the utterances belonging to an interaction, so the distances in the original space are preserved.

We use the t-distributed stochastic neighbor embedding (t-SNE) algorithm [26], a nonlinear dimensionality reduction method that preserves the neighborhood information from an original high-dimensional space in the projected low-dimensional space, to visualize the difference between AMS-normalized and CMS-normalized features. In Figure 1, CMS-normalized non-speech features are surrounded by the background speech features and the desired and background speech overlap. With AMS-normalized features, depicted in Figure 2, we observe that the non-speech, desired, and background (BG) speech features are more separable, which is consistent with the assertion that AMS better preserves speaker differences.

2.2. Encoder-Decoder Model

Although AMS has many advantages over CMS and has been shown to work well in the related task of desired speech detection [24], it can be suboptimal in two aspects: (a) the representation of the desired speech, defined as the mean of the anchor word features, is hand-crafted and not learned; (b) its interaction with the acoustic features of an utterance is also predefined (as subtraction), instead of being learned. With the encoder-decoder model, we extend AMS by learning both the representation and its interaction with the acoustic features from the training data.

Inspired by the encoder-decoder framework used in machine translation community [27], we explored a variation of it for desired speech detection in [24], as depicted in Figure 3. The encoder network in Figure 3, which is a single-layer LSTM model consuming a variable length sequence of features from the anchor word segment, generates an embedding of the desired speech. This embedding is appended to the acoustic feature vector fed into a decoder DNN, which is trained to predict senone posteriors for the frame. Though decoders are usually recurrent networks in an encoder-decoder framework, here we use a feed-forward DNN for simplicity. The parameters are jointly optimized by minimizing the cross-entropy objective function via mini-batch stochastic gradient descent.

3. Experimental Setup

We perform experiments on an in-house dataset consisting of anonymized, hand-transcribed utterances from the Amazon Echo. All the utterances begin with the same anchor word. A keyword spotter algorithm such as [28] was used to automati-

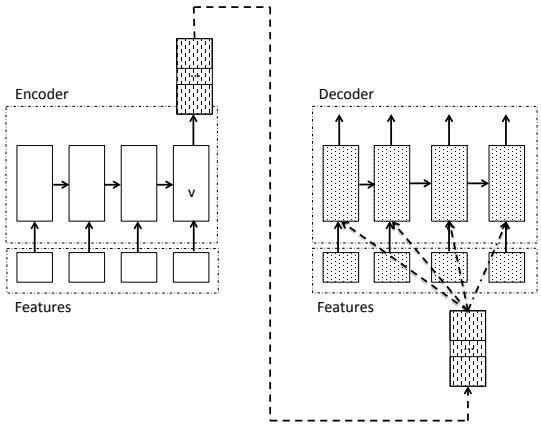

Figure 3: *Encoder-Decoder architecture for anchored speech recognition. The input of the encoder is the anchor word segment spoken by the desired speaker. An embedding vector of the desired speaker, i.e. last output of the encoder 'v', is then extracted and fed in the decoder alone with the acoustic features for speech recognition [24].*

cally provide the timing information for the anchor words. The training and test data amount to 1,200 and 10 hours, respectively. Speakers in the test set were excluded from the training set.

The following three systems were built using 64-dimensional log filterbank energy (LFBE) features extracted every 10ms with a window size of 25ms: (a) a baseline DNN AM using CMS-normalized features, (b) a AMS DNN AM using AMS-normalized features, and (c) an encoder-decoder (EncDec) AM using AMS-normalized features. All DNN AMs had 8 hidden layers, 1664 neurons per layer, and an output size of 3101. The DNN inputs were whitened with the corresponding global mean and variances and spliced with +/- 8 frames of context. The decoder AM in the EncDec network was initialized with the AMS DNN and the decoder was initialized with random values. An LSTM encoder layer of size 32 was used. The encoder was fed AMS-normalized features from the anchor word segment. The output of the last frame from the encoder was then combined with the first layer of the DNN via an affine transform. All the networks were trained with the cross-entropy objective using a distributed trainer [29].

4. Results and Discussion

In this section, we will present ASR results on the overall dataset, on subsets of utterances with and without background and media speech, and on subsets grouped by their signal-to-noise ratio.

4.1. Overall Results

The first row in Table 1 shows the performance of the three systems on the test dataset. The improvements predominantly came from reducing insertions. This result supports our hypothesis that the anchor-based systems should be more robust to interfering background speech, which commonly manifest as insertions. We hypothesize that the superior performance of the EncDec system is due to learning a more optimal anchor word embedding directly from the data versus AMS' mean subtraction.

Table 1: *Comparison of the three systems on the overall test set as well as in different background conditions including desired speech (DS), multimedia speech (MS), and background speech (BG). The word error rate, substitution, and deletion values are all normalized by the baseline WER, and the word error rate reduction (WERR) are evaluated as relative WER improvement over the to the baseline CMS system. For example, if the word error rate for the baseline is 10% and the substitution rate is 5%, then the normalized values would be 1.000 and 0.500.*

	WER	**sub**	**ins**	**del**	**WERR**
Overall					
CMS	1.000	0.427	0.424	0.148	—
AMS	0.935	0.430	0.351	0.154	+6.5
EncDec	0.908	0.421	0.334	0.152	+9.2
DS Only					
CMS	1.000	0.624	0.151	0.224	—
AMS	1.019	0.648	0.149	0.222	-1.9
EncDec	0.982	0.625	0.137	0.219	+1.8
DS+MS					
CMS	1.000	0.396	0.483	0.121	—
AMS	0.832	0.396	0.307	0.129	+16.8
EncDec	0.819	0.395	0.290	0.134	+18.1
DS+BG					
CMS	1.000	0.259	0.681	0.060	—
AMS	0.847	0.241	0.548	0.058	+15.3
EncDec	0.824	0.239	0.522	0.063	+17.6
DS+BG+MS					
CMS	1.000	0.276	0.690	0.035	—
AMS	0.724	0.190	0.483	0.052	+27.6
EncDec	0.647	0.198	0.397	0.052	+35.3

tion.

4.2. Background speech analysis

To better understand the performance of the three systems, we slice the test set into four different non-overlapping subsets: (a) utterances with desired speech (DS) only, (b) utterances with multimedia speech[1] (MS), (c) utterances with background speech (BG), and (d) utterances with both multimedia and background speech for analysis. The rest of the four rows in Table 1 show performance of the three systems in these four conditions.

4.2.1. Desired speech only

We will first look at performance on utterances containing only desired speech. Since CMS is continually adapting to better recognize all speech, this is the condition in which CMS is best suited, compared to the other conditions. The three systems had very similar WERs, which we count as a success; although the AMS and EncDec models were designed for robustness to BG speech, they do not perform significantly worse in the desired speech-only condition. The EncDec system even has a 1.8% relative improvement over the baseline CMS system.

[1]Multimedia speech means that a television, radio, or other media device is playing back speech in the background. Music or other sounds are not classified as multimedia speech and are included in the desired speech-only subset.

4.2.2. Multimedia speech

Next, we look at the utterances with multimedia speech. We observed that the multimedia device played speech that was more compressed (i.e. less variation in volume) than natural speech and also contained a significant amount of non-speech sounds. For example, commercials commonly contain overlapping speech, music, and sound effects. This leads to a higher relative level of substitution errors in this condition compared with utterances containing background speech, due to more interfering speech/noise overlapping with the desired speech. While the systems had similar substitution and deletion errors as the baseline, the insertion errors of the two systems were significantly reduced.

4.2.3. Background speech

For utterances containing desired and background speech, the AMS and EncDec systems still outperformed the baseline CMS significantly. The two approaches successfully assisted acoustic models to be more robust to background speech, greatly reducing the number of insertion errors.

4.2.4. Multimedia and background speech

Finally, in utterances containing both multimedia and background speech, AMS and EncDec consistently outperform the baseline CMS system. It is interesting to note that in this condition, which is the most complex of the above scenarios, both the anchored ASR systems had a significant reduction of substitution errors, and the EncDec system achieved a much lower number of insertion errors than the AMS system. The AMS and EncDec systems achieved 27.6% and 35.3% relative WER improvement over the baseline CMS, respectively. This shows that the two proposed approaches are able to handle these difficult scenarios containing interfering speech.

4.3. SNR analysis

Next, we slice the data by signal-to-noise ratio (SNR) in Table 2 to see the systems' robustness to noise. In order to calculate the SNR on the utterances, we first classify each frame as desired speech, human background speech, and noise via forced alignment. Frames containing multimedia speech are classified as non-speech. We estimate the SNR via $\frac{DS-NS}{NS}$, where DS and NS are the average energies of the desired speech and non-speech frames, respectively. The noise is not stationary, but observed to be uncorrelated enough with the desired speech that we can model them as uncorrelated. Thus, in order to calculate the average speech energy from the frames containing speech, we subtract the estimated average noise. We acknowledge that this is an imperfect measurement of SNR, but believe it is useful enough for use in comparing ASR performance of the systems in different noise levels.

We group the SNR values into three conditions: 20-30 dB, 10-20 dB, and 0-10 dB. We have discarded extremely high and low SNR utterances. First we will compare across different SNR conditions. When comparing the AMS and EncDec systems to the baseline, we see a 4.7% and 7.5% improvement across 20-30 dB, 9.3% and 12.5% improvement across 10-20 dB, and 3.9% and 4.6% improvement across 0-10 dB. We hypothesize the following reasons. In the 20-30 dB SNR condition, the anchored ASR systems have fewer insertions and more substitutions than the baseline. This is due to CMS adapting to better match the desired speech instead of interfering speech or noise, since the noise level is relatively low. In the 10-20 dB SNR condition, the improvements with the anchored ASR systems are most significant. This is because the interfering speech is loud enough to start being recognized by the continually-adapting CMS model, but still different enough to be properly rejected by the AMS and EncDec models. In the 0-10 dB SNR condition, the noise and interfering speech levels are so high that they are more similar to the desired speech features. Also, in the case that the interfering speech or noise overlaps with the anchor word segment, it is more likely to corrupt the anchor word features and lead to degraded performance with the anchored ASR systems. Finally, in comparing the AMS and EncDec systems, we see that the EncDec system has better WER, substitution, insertion, and deletions than the AMS system, the only exceptions being substitutions and insertions are both slightly higher on the 0-10 dB SNR condition.

Table 2: *Comparison of the three systems in different signal-to-noise (SNR) conditions. The word error rate, substitution, and deletion values are all normalized by the baseline WER, and the word error rate reduction (WERR) are evaluated as relative WER improvement over the to the baseline CMS system.*

	WER	**sub**	**ins**	**del**	**WERR**
20-30 dB					
CMS	1.000	0.505	0.304	0.191	—
AMS	0.953	0.530	0.218	0.206	+4.7
EncDec	0.925	0.526	0.197	0.200	+7.5
10-20 dB					
CMS	1.000	0.457	0.385	0.159	—
AMS	0.907	0.446	0.299	0.161	+9.3
EncDec	0.875	0.435	0.283	0.157	+12.5
0-10 dB					
CMS	1.000	0.374	0.498	0.128	—
AMS	0.961	0.377	0.445	0.139	+3.9
EncDec	0.954	0.354	0.454	0.145	+4.6

We also looked at signal-to-interferer ratio (SIR), which we estimated as $\frac{DS-NS}{BS-NS}$. Since SIR is only defined for utterances containing human background speech, some of the SIR slices have small utterance counts, so we did not include all the results in a table. However, we see that as SIR increases, the relative improvement in WER increases with both anchored ASR systems: 6% at 0-10 dB, 21% at 10-20 dB, 38% at 20-30 dB, and 50% at 30-40 dB.

5. Conclusions

In this paper, we investigated both feature-based and model-based methods for improving AM performance in the presence of interfering speech. For the anchored mean subtraction approach, we use the mean estimated from the anchor word to normalize the speech features in the utterance. For the encoder-decoder approach, we use an encoder to create an embedding of the anchor word, which is designed to capture the characteristics of that speaker. Results on our in-house dataset reveal that the anchored ASR methods outperform the baseline CMS method overall and on all the different interfering speech and SNR conditions, with up to a 35% relative WER improvement. In all conditions, the encoder-decoder network performed the best, demonstrating that the more-powerful network can learn better features than the hand-crafted AMS features.

6. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The Microsoft 2016 conversational speech recognition system," in *International Conference on Acoustics, Speech and Signal Proceedings (ICASSP)*. IEEE, 2017, pp. 5255–5259.

[2] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2013, pp. 1–4.

[3] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[4] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2013, pp. 162–167.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[6] M. Harper, "The automatic speech recogition in reverberant environments (ASpIRE) challenge," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 547–554.

[7] P. Bell, M. J. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester *et al.*, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 687–693.

[8] T. G. Stockham, T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 678–692, 1975.

[9] R. Schwartz, T. Anastasakos, F. Kubala, J. Makhoul, L. Nguyen, and G. Zavaliagkos, "Comparative experiments on large vocabulary speech recognition," in *Workshop on Human Language Technology*. Association for Computational Linguistics, 1993, pp. 75–80.

[10] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[11] O. Viikki and K. Laurila, "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization," in *Robust Speech Recognition for Unknown Communication Channels*, 1997.

[12] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[13] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. IEEE, 1996, pp. 733–736.

[14] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.

[15] X. Cui and A. Alwan, "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1161–1172, 2005.

[16] J. Droppo and A. Acero, "Environmental robustness," in *Springer handbook of speech processing*. Springer, 2008, pp. 653–680.

[17] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7398–7402.

[18] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE, 2014, pp. 172–176.

[19] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 436–443.

[20] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin *et al.*, "Meetings about meetings: research at ICSI on speech in multiparty conversations," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4. IEEE, 2003, pp. IV–740.

[21] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 84–91, 2005.

[22] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," in *International Conference on Spoken Language Processing (ICSLP)*, no. LIDIAP-CONF-2006-007, 2006.

[23] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grézl, A. El Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[24] R. Maas, S. H. K. Parthasarathi, B. King, R. Huang, and B. Hoffmeister, "Anchored speech detection," in *17th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2016, pp. 2963–2967.

[25] S. Tibrewala and H. Hermansky, "Multi-band and adaptation approaches to robust speech recognition," in *Eurospeech*. ISCA, 1997.

[26] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[27] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2014.

[28] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, p. 4087–4091.

[29] N. Strom, "Scalable distributed DNN training using commodity GPU cloud computing," in *16th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2015.