Interspeech 2018
2-6 September 2018, Hyderabad

Cross-Corpora Convolutional Deep Neural Network Dereverberation Preprocessing for Speaker Verification and Speech Enhancement

Peter Guzewich[1], *Stephen Zahorian*[1], *Xiao Chen*[1], *Hao Zhang*[1]

[1]Department of Electrical and Computer Engineering, Binghamton University, NY, USA
[peter.guzewich,zahorian,xchen49,hzhang20]@binghamton.edu

Abstract

Deep neural network (DNN) dereverberation preprocessing has been shown to be a viable strategy for speech enhancement and increasing the accuracy of automatic speech recognition and automatic speaker verification. In this paper, an improved DNN technique based on convolutional neural networks is presented and compared to existing methods for speech enhancement and speaker verification in the presence of reverberation. This new technique is first shown to enhance speech quality as compared to other existing methods. Then, a more thorough set of experiments is presented that assesses cross-corpora speaker verification performance on data that contains real reverberation and noise. A discussion of the applicability and generalizability of such techniques is given.

Index Terms: dereverberation, convolutional deep neural networks, speech quality, speaker verification

1. Introduction

Deep neural networks (DNNs) have seen a surge of interest in the research community, having been successfully applied to a large number of different tasks including automatic speech recognition (ASR) and speaker identification (SID). This explosion in research is fueled in part by the prevalence of so-called “big data” in modern systems. Because deep learning is a data driven approach, it is only with the help of modern large databases that networks can be fine-tuned to a degree not possible in the past. However, in many cases, databases are not specifically suited to the task at hand or it isn’t possible to leverage them in a given situation. Therefore, the “not enough data” problem is still a real possibility, particularly if the training data is too dissimilar from the test data.

While current state-of-the-art speech processing systems perform well, they often degrade with speech from noisy or reverberant environments. Many preprocessing techniques have been proposed to deal with these situations (e.g. [1] [2] [3] [4]), with many also including deep learning (e.g. [5] [6] [7] [8] [9] [10]). In this paper, recent deep neural network dereverberation preprocessing techniques are addressed. An improved technique based on convolutional neural networks is proposed and its performance is experimentally shown to be better than other methods. Finally, a detailed analysis of performance on real data containing real reverberation and noise, as opposed to simulated, is presented.

2. Background

2.1. Dereverberation with DNNs

Sound waves traveling in a reverberant environment reflect off of walls and objects. The overlapping waves produce temporal/spectral smearing which reduces the intelligibility of speech [11] and degrades performance for both ASR and SID. Recent work has examined the potential of preprocessing speech waveforms with deep neural networks for speech enhancement and to increase performance of other speech processing tasks [7] [8] [9] [10]. The general paradigm for these methods involves mapping spectral representations of individual frames of reverberant speech to estimates of the spectrum of corresponding clean speech frames. The details of these works differ, but all follow this general principle: using a multilayer feedforward deep neural network to enhance frames of speech. In each of these, speech quality and intelligibility were used as metrics for speech enhancement performance. In [8], Han et al. also showed improved ASR performance. In [10], the method improved speaker verification performance. These works showed the technique was beneficial for these other speech tasks, despite not being designed to discriminate between phonemes or speaker specific characteristics.

2.2. Reverberant and noisy data

An important detail of the aforementioned works is that they all used only artificially reverberant speech. Due to the nature of the approach, training the networks requires having perfectly matched sets of clean and reverberant data. There are databases that could meet this need to some degree, but it is generally much easier to create the data artificially. In each of those works, the databases were generated by corrupting clean speech waveforms by convolving them with room impulse responses (RIRs) to produce reverberant versions. While this is an easy way to create data, good results on tests performed on this data don’t necessarily extend to real reverberation.

To further explore the issue of real reverberation, this study also makes use of the Multiroom8 corpus provided by the Air Force Research Lab (AFRL) in Rome, NY. This database is fairly small, containing 807 spontaneous and prompted speech utterances from 52 speakers totaling about 40 hours of speech recorded in 4 different sized rooms, 3 containing 6 microphones each. Multiple microphones were used during the recordings, each placed in a different location to provide signals with varying degrees of signal-to-noise ratio (SNR) and reverberation. This dataset was used to allow for a loose comparison with the previous work of [4]. More specific details on the database and setup can be found in that work. The experimental results given later in section 4 are not directly comparable to those given in [4], but a direct comparison between the methods is made in section 4.

As the waveforms used were recorded in common environments (e.g. conference rooms), they contain unwanted noise, such as humming sounds from the building’s heating system, as well as reverberation. This “real world” data therefore adds a complication to the work, that of simultaneous reverberation and noise mitigation. It has been

10.21437/Interspeech.2018-2238

suggested that these items are best left separated [12], but in this paper, they are considered jointly, although somewhat indirectly.

2.3. Convolutional deep neural networks

Convolutional neural networks became quite well-known in the last decade for their incredible performance for image processing and classification. Perhaps the most famous paper on the subject introduced a convolutional network known as AlexNet [13]. A convolutional network layer includes a number of filters that are each used to process the input. The goal is to train a series of two-dimensional convolutional filters to produce activation maps that hold important information. The key point behind these two-dimensional filters is that they can be trained to search for patterns that exist in the spatial relationship of adjacent input values. As mentioned, these networks gained notoriety for image processing because they could be used to pick out things like shapes in images and therefore have been found to be useful for image classification.

Speech waveforms are one-dimensional, but they are often represented in a time-frequency plane commonly known as a spectrogram. By transforming the signal into a spectral representation, which is nearly always the first step for speech processing tasks, a speech signal can therefore be transparently inserted into the paradigm of convolutional networks. The spectral representation of speech exhibits frequency and temporal correlations, so the imposed structure of a convolutional network can therefore be expected to be well suited to the task of extracting important information. Compared with feedforward networks previously used, this structure can be expected to be a useful guide for the network during the training process.

2.4. Performance metrics

Performance of the dereverberation and denoising networks was first judged on the basis of speech quality scores. However, the ultimate goal of the work is to improve performance for SID tasks, so experiments were also done to examine performance for speaker verification. Specifically, perceptual evaluation of speech quality (PESQ) [14] and short-time objective intelligibility (STOI) [15] were used as the basis for speech enhancement performance. Then, equal error rates (EERs) for speaker verification were used to quantify SID task performance. Speaker verification and speaker identification are quite similar tasks. Therefore reduced EERs are a good indication that SID performance would improve as well.

3. Convolutional neural network strategy for improved dereverberation

The proposed technique is an improvement on our previous work [10] with the main framework of the processing being similar, but with a few key modifications. The most notable change is that the feedforward deep neural network was replaced with one based on a convolutional structure.

As mentioned above, training the network requires matched sets of reverberant and clean data. For all artificially created data, RIRs generated with the improved image source method [16] for a range of T60 values were convolved with clean waveforms to produce the reverberant versions. For each RIR, one reverberant copy of the data was produced. Training waveforms were again time-aligned to the point of maximum cross-correlation, but the amplitudes were not directly scaled to be equal. Instead, the waveforms were scaled so that the variances of their amplitudes were all equal. This strategy boosts performance by better accounting for the dynamic nature of speech across different speakers and channels.

After time domain alignment processing, the waveforms were transformed into log-magnitude spectral values via a fast Fourier transform (FFT) of size 512, thus creating 257 magnitude values for each frame. This size, rather than a larger one, was chosen as a compromise to computational complexity at training time because convolutional network layers generally take longer to train than feedforward layers. The log-magnitude spectrum was then normalized using the common mean and variance normalization (MVN) strategy on a per-frequency-bin basis. These values were then passed through the network (several context frames of input, one frame of output).

3.1. Proposed network structure

The network structure is a familiar one, based on a well-known work in the computer vision field, VGGNet [17]. The VGGNet network utilizes a large number of small convolutional filters in succession to emulate the capability of larger network layers. The key insight of this strategy is that small filters (e.g. 3x3) do not envelope much area, but if combined in succession can have a similar input space as a larger single layer. The advantage of this method is that the layers are quicker to train as they contain fewer parameters. The pseudo-code representing the network structure is shown in Figure 1. The proposed network includes a total of nine convolutional layers with noted number/size (e.g. 32, 3:3) filters, four pooling layers, and two final feedforward layers. Between all layers is a rectified linear activation function.

```
ConvolutionalLayer {32 filters, kernel (3:3)} : ReLU
ConvolutionalLayer {32 filters, kernel (3:3)} : ReLU
MaxPoolingLayer    {kernel (2:2), stride=(2:1)}
ConvolutionalLayer {64 filters, kernel (3:3)} : ReLU
ConvolutionalLayer {64 filters, kernel (3:3)} : ReLU
MaxPoolingLayer    {kernel (2:1), stride=(2:1)}
ConvolutionalLayer {128 filters, kernel (3:3)} : ReLU
ConvolutionalLayer {128 filters, kernel (3:3)} : ReLU
MaxPoolingLayer    {kernel (2:1), stride=(2:2)}
ConvolutionalLayer {256 filters, kernel (3:3)} : ReLU
ConvolutionalLayer {256 filters, kernel (3:3)} : ReLU
ConvolutionalLayer {256 filters, kernel (3:3)} : ReLU
MaxPoolingLayer    {kernel (2:1), stride=(2:2)}
FullConnectlayer   {2048} : ReLU : Dropout
FullConnectlayer   {2048} : ReLU : Dropout
```

Figure 1: *CNTK pseudo-code for proposed network*

The networks were trained with the Microsoft Research (MSR) Cognitive Toolkit (CNTK) [18] to minimize square error between the estimated spectrum and clean spectrum. Parallel processing was done with an NVidia GTX TITAN GPU card, but training still took several days, depending on the dataset. A forward and backward context of 7 frames was used for a 257x15 dimensional input. As discussed in [9] and [10], changing the frame context window affected performance differently for different T60 times (higher T60 benefits from more context). A context of 15 total frames was chosen as a compromise. The network learning rate was initialized at 0.0001 per batch and decayed by half every few epochs. The minibatch size was 128, the dropout rate was 0.5, and the momentum term was 0.8.

4. Experiments

A series of experiments were performed to demonstrate the speech enhancement and speaker identification potential of the proposed technique. First, the technique was evaluated on the basis of speech quality and intelligibility using artificially reverberant versions of the TIMIT [19] database. Then, to test for SID potential in reverberant and noisy conditions, experiments were performed using the Multiroom8 corpus and artificially reverberant versions of the phone conversation portion of the Mixer 6 database [20].

4.1. Speech Enhancement

To directly compare with existing speech enhancement results, the same procedure was followed as in [10], which itself was designed to mimic the setup of [9]. Briefly, 10 artificial RIRs were generated for a range of T60 values (0.1s to 1s in 0.1s increments) and convolved with the entire TIMIT training set of 4620 utterances for a total of approximately 40 hours of reverberant speech. In [10], the training set included a copy of the clean data, but in this experiment it was left out. This point is addressed later in the paper. The test set was a random selection of 100 utterances from the TIMIT test set convolved with 20 newly generated RIRs with T60s ranging from 0.05s to 1s in 0.05s increments. Using the entire TIMIT test set produced similar results.

The network (Proposed) was trained and used to process test waveforms. The waveforms were also processed with Blind Spectral Weighting (BSW) [4], Temporal Masking and Thresholding (TMT) [3], an implementation of Wu's network from [9] (Wu), and by our previous network (Guzewich) in [10] for comparison. Average PESQ scores were computed for all cases including the ideal (Ideal) and unprocessed waveforms (Reverb) and shown in Figure 2. Figure 3 shows a sample reverberant spectrogram and its processed counterpart.

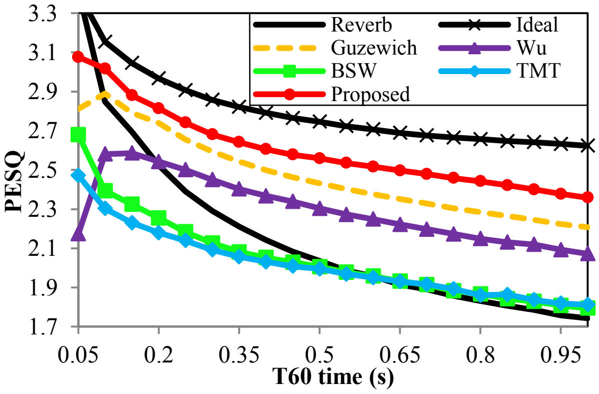

Figure 2 : *Average PESQ scores for Proposed, BSW, TMT, Guzewich, Wu, and baselines*

As is shown in the figures, the proposed method improves speech quality scores much more than the other methods. An interesting result from this experiment is the performance on fairly clean data (T60 < 0.2s). As mentioned, this proposed network was not trained with any clean data input, which was an important part of the previous work [10] (Figure 2, labeled Guzewich). Despite the use of "less" training data, the proposed network still outperformed the network from previous work for fairly clean data and there was no "drop-off" in performance in that region as with the other two neural network methods. The proposed method produced scores remarkably close to the ideal baseline result on the top of the graph, which corresponds to perfect restoration of the magnitude spectrum combined with the reverberant phase.

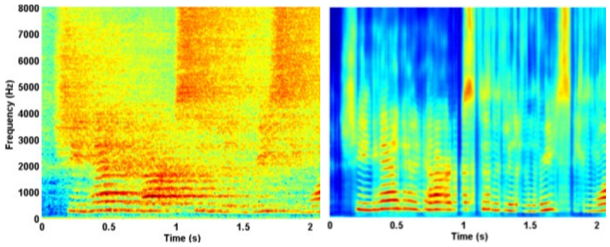

Figure 3 : *An example reverberant TIMIT segment (left) and DNN processed (right)*

All the DNN processed waveforms were also evaluated with STOI, and average scores are shown in Figure 4.

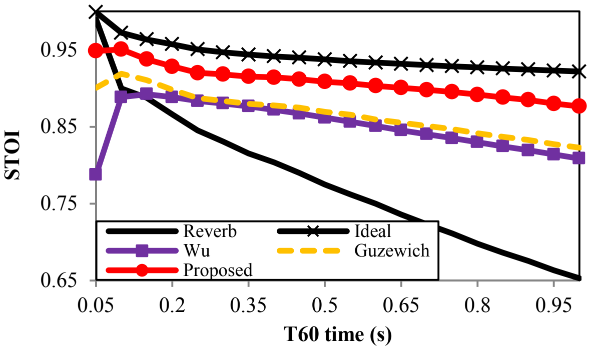

Figure 4: *Average STOI scores for Proposed, Guzewich, Wu, and baselines*

4.2. Speaker verification

Ideally speaking, a processed sentence from the proposed technique should appear as though it is a perfectly clean one, but although the proposed technique improves speech quality and intelligibility scores, it is not certain to improve performance for other speech processing tasks, especially for data containing real reverberation and noise. To assess potential for SID tasks, several speaker verification experiments were performed using the proposed technique without any further optimization for the SID task. These experiments used data with real reverberation and real noise from the Multiroom8 corpus. A portion of the Mixer 6 database was also used which contained ~14 hours of telephone conversation speech from 594 speakers (292 males, 302 female). This Mixer 6 data was chosen instead of TIMIT because it has the same bandwidth as Multiroom8.

All speaker verification experiments were done using the Alize [21] iVector system with probabilistic linear discriminant analysis (PLDA) scoring. They used a 1024 mixture universal background model, iVector dimension of 200, and a PLDA Eigenvoice and Eigenchannel dimension of 100 and 50, respectively. All of these models were trained using the ~14 hours of original Mixer 6 data. The enrollment and test data was some portion of the Multiroom8 corpus as noted for each experiment. There were 7 enrollment and test configurations which were based on those used in [4]. Tables 1 and 2 contain these conditions in the first columns. The notation is enrollment-test, so for example, Enroll-Sm4 means enrollment data is from the enrollment (conference) room microphone and test data is from the 4th microphone from the small room. There were 35 speakers that were present in every configuration, but as is noted later in this paper, when the dataset is split, speakers are split up (e.g. 17 speakers in one half, 18 in the other) so no common speakers occur between groups. The features used for speaker verification were Mel-Frequency Cepstral Coefficients (MFCCs) [22] with a 25ms frame length and 10ms frame shift. We use the first 13 terms (excluding C0) with delta and delta-delta terms for a total of 39 features. All speaker verification results are shown alongside baseline MFCCs, computed on the unprocessed data, and MFCCs computed from data processed by BSW, for comparison.

The first set of results shows the technique was beneficial for the task of speaker verification, specifically with data containing real noise and reverberation. Two DNNs were trained, in round-robin fashion, using each half of the Multiroom8 corpus. Then, each network was used to process the opposite half of the data to be used in a verification experiment. This condition is labeled RR-Mult. Network training was possible by assuming the signals from the microphones situated closest to the speakers as the clean signals, even though these signals were not perfectly clean. It is likely that a dataset created with this DNN technique in mind would allow for better training.

The next results show cross-corpora performance in the context of the "not enough data" problem. Another network was trained using only the Mixer 6 data, which was again corrupted 10 times using generated RIRs with the same strategy used for the training data in section 4.1. This network was used to process the Multiroom8 corpus for verification and this condition is labeled MX6. The results are shown in Table 1, which gives error rates based on the full dataset.

Table 1: *EERs for entire Multiroom8 corpus*

	MFCC	RR-Mult	MX6	BSW
Enroll-Sm4	11.43	7.73	18.99	12.61
Enroll-Sm6	14.87	5.71	15.04	15.21
Lg4-Med5	17.06	14.29	22.86	20.25
Lg5-Sm4	14.29	11.85	17.14	12.35
Med3-Sm3	6.72	8.57	10.84	8.57
Med5-Sm5	5.71	5.71	5.38	2.86
Sm4-Lg5	14.29	11.85	17.14	12.35
Average	12.05	9.39	15.34	12.03

As shown in Table 1, the proposed DNN technique was able to reduce average speaker verification error rates by 22% compared with the baseline on data that contains real reverberation and noise. This is a key result, because all the other DNN methods discussed have never been tested on such data. This is also significant because there was no explicit optimization of the technique for this speaker verification task. The technique was tuned for speech enhancement, but is still capable of improving performance for SID related tasks. Another result from this table is that, unsurprisingly, using acoustically similar data between testing and training improved performance (RR-Mult is better than MX6).

The trained network from MX6 was then used as the base for transfer learning, whereby the trained network parameters were frozen except for the final feedforward layers. The pre-trained network was then trained again using ¼ of the Multiroom8 data, but only the final feedforward layers were actually learning. This condition is labeled MX6-Mult-Q1. For comparison, a new network was trained using only the same ¼ of the Multiroom8 data and this condition is labeled Mult-Q1. Both networks were used to process the remaining ¾ of the Multiroom8 data to be used for verification. Table 2 shows these results, as error rates for ¾ of Multiroom8.

Table 2: *EERs for ¾ Multiroom8 corpus*

	MFCC	MX6-Mult-Q1	Mult-Q1	BSW
Enroll-Sm4	10.11	8.55	11.11	11.40
Enroll-Sm6	14.39	7.41	6.55	7.41
Lg4-Med5	18.38	15.81	16.10	17.09
Lg5-Sm4	11.11	11.11	11.25	11.11
Med3-Sm3	7.41	6.70	11.11	7.55
Med5-Sm5	4.84	7.26	7.41	3.70
Sm4-Lg5	11.11	11.11	11.25	11.11
Average	11.05	9.71	10.68	9.91

As shown in Table 2, the proposed technique lowered average speaker verification error rates compared to the baseline. These results are more interesting, however, due to the restricted data size. Here, the emphasis of the experiment was on the "not enough data" problem. When training the network using only a small portion of the acoustically similar data from Multiroom8, the error is reduced only marginally (3% reduction). BSW performs better than this. However, with the use of transfer learning on Mixer 6 data supplemented by Multiroom8, the error is reduced by 12% compared with the baseline. This result supports the intuitive conclusion that "there's no data like more data." Generalization is a key feature of a successful neural network, so the more data that can be obtained for training, the better.

5. Discussion

In this paper, an improved DNN dereverberation technique based on convolutional neural networks was proposed. This technique was developed primarily for improved speech enhancement and was experimentally shown to improve speech quality scores more than existing methods. Then, the technique was extended to the task of speaker verification despite not being specifically developed for that purpose. In a series of experiments, the method was shown to be capable of reducing error rates for speaker verification on data that contained real reverberation and noise which had not been shown by the compared works. The experiments were also designed to examine the cross-corpora performance of the technique. A key takeaway is a common technique known as transfer learning can help boost performance in cases where there isn't enough relevant data. The results support the logical conclusion that neural network performance is likely to improve with more training data, especially if the technique is further optimized for the task of speaker verification. To that end, our current work includes exploring modifications to the technique in order to specifically improve SID performance with and without adequate data.

6. Acknowledgement and Disclaimer

This material is based on research sponsored by the Air Force Research Laboratory under contract FA8750-16-C-0292 awarded to Minerva Systems & Technologies, LLC and Binghamton University.

Approved for public release: 88ABW-2018-3325

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the Air Force Research Laboratory or the U.S. Government.

7. References

[1] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[2] A. Jukic and S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *ICASSP*, 2014.

[3] C. Kim, K. K. Chin, M. Bacchiani and R. M. Stern, "Robust speech recognition using temporal masking and threshold algorithm," in *Interspeech*, Singapore, 2014.

[4] S. O. Sadjadi and J. H. Hansen, "Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014.

[5] M. Mimura, S. Sakai and T. Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders," in *Proceedings of REVERB Challenge Workshop*, 2014.

[6] L. X., T. Y., S. Matsuda and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech*, 2013.

[7] K. Han, Y. Wang and D. Wang, "Learning Spectral Mapping for Speech Dereverberation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014.

[8] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks and T. Zhang, "Learning Spectral Mapping for Speech Dereverberation and Denoising," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015.

[9] B. Wu, K. Li, M. L. Yang and C.-H. Lee, "A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[10] P. Guzewich and S. Zahorian, "Improving Speaker Verification for Reverberant Conditions using Deep Neural Network Dereverberation Processing," in *Interspeech*, Stockholm, 2017.

[11] P. Assmann and A. Summerfield, "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, New York, Springer, 2004, pp. 231-308.

[12] Y. Zhao, D. Wang, I. Merks and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *ICASSP*, 2016.

[13] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, 2012.

[14] A. Warzybok, I. Kodrasi, J. O. Jungmann, E. A. P. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollomier and S. Goetze, "Subjective Speech Quality and Speech Intelligibility Evaluation of Single-Channel Dereverberation Algorithms," in *IWAENC*, Antibes, France, 2014.

[15] C. Taal, R. Hendriks, R. Heusdens and J. Jensen, "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech," in *ICASSP*, 2010.

[16] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Acoustical Society of America,* vol. 124, pp. 269-277, 2008.

[17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR*, 2014.

[18] A. Agarwal, E. Akchurin, C. Basoglu, G. Chen, S. Cyphers, J. Droppo, A. Eversole, B. Guenter, M. Hillebrand, X. Huang, Z. Huang, V. Ivanov, A. Kamenev, P. Kranen, O. Kuchaiev, W. Manousek and A. May, "An Introduction to Computational Networks and the Computational Network Toolkit," Microsoft Technical Report, 2014.

[19] W. M. Fisher, G. R. Doddington and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," 1986.

[20] L. Brandschain, D. Graff and K. Walker, "Mixer 6 Speech LDC2013S03," Linguistic Data Consortium, 2013.

[21] A. Larcher, J. Bonastre and H. Li, "ALIZE 3.0 - Open-source platform for speaker recognition," IEEE SLTC Newsletter, 2013.

[22] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1980.