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

Improving Speaker Verification for Reverberant Conditions with Deep Neural Network Dereverberation Processing

Peter Guzewich[1], *Stephen Zahorian*[1]

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

Abstract

We present an improved method for training Deep Neural Networks for dereverberation and show that it can improve performance for the speech processing tasks of speaker verification and speech enhancement. We replicate recently proposed methods for dereverberation using Deep Neural Networks and present our improved method, highlighting important aspects that influence performance. We then experimentally evaluate the capabilities and limitations of the method with respect to speech quality and speaker verification to show that ours achieves better performance than other proposed methods.

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

1. Introduction

Modern speech processing tasks, such as automatic speech and speaker recognition, can be performed with high accuracy in clean, controlled conditions. However, there continues to be a need for and much interest in improving performance for speech contaminated by noise, channel effects, and reverberation. Specifically, reverberation is known to degrade performance of both speech and speaker recognition systems. This is commonly attributed to the spectral/temporal smearing caused by the addition of late reflected components of the speech [1]. The late-arriving reflections cause a masking effect and fill in the otherwise silent portions of the speech.

Many front-end signal processing methods have been described in the literature to cope with these effects (for example, [2] [3] [4] [5] [6] [7]). These methods seek to improve the raw signal or information extracted from it to mitigate the degrading effects of reverberation. Back-end modeling/ decision methods [8] [9] seek improvement from changes in the feature modeling or the recognizer. Recently, deep neural networks (DNNs) have been the subject of much study for the task of front-end enhancement in the feature or signal domain [8] [10] [11] [12] [13] [14].

In this paper, we address recently reported methods [11] [12] [14] for DNN based enhancement of reverberant speech and illustrate some important details, which have a large effect on performance, when using the method. We discuss the capabilities and limitations of the paradigm as they relate to speech quality. We also introduce modifications that improve speaker verification and speech quality for various conditions. The primary contributions of this work are to document the complex process of using DNNs for dereverberation and show how proposed modifications help improve speech quality and performance for speaker verification. Note that speaker verification is similar to speaker identification (SID), so presumably techniques which improve verification accuracy will also improve SID accuracy.

The rest of this paper is organized as follows. Section 2 provides the background for the method. Section 3 gives details of our implementation and modifications. Section 4 presents the experimental setup and results for speech quality and speaker verification. Finally, Section 5 draws some conclusions.

2. Background

2.1. Dereverberation using deep neural networks

When speech is produced in a reflective environment, the reflected wave fronts combine with the direct sound. This has the effect of degrading performance of speech processing tasks because of temporal and spectral smearing. Recently reported work [11] [12] [14] shows promising results for improving the quality of speech recorded in reverberant conditions using a DNN. These works represent the first examples in the literature of this type of processing of the signal. There are not yet any reported studies which apply this method to SID tasks.

In [11], Han et al trained a DNN to map frames of artificially reverberant speech (161 gammatone filterbank magnitude values from speech generated by convolving artificial room impulse responses (RIRs) with clean data) to the corresponding frames of clean speech. They first used short time Fourier transform (STFT) spectral magnitudes, but achieved better performance with filterbank outputs. However, it has been shown that a network is capable of learning an adaptable pseudo-filterbank response [15], suggesting there is nothing necessarily gained by using a pre-specified filterbank. In their follow-up work [12], Han et al switched to using log magnitude STFT spectrum values. They used a network with 3 hidden layers with rectified linear activations and an output layer with a sigmoid activation function. They mapped the target features into the unit range of [0,1] while mapping the input features to zero mean and unit variance. The output of the trained network was the estimated clean speech log magnitude spectrogram, which was rescaled from the unit range and recombined with the phase from the original reverberant waveform via overlap-and-add with optional post-processing [16] to produce an estimate of the anechoic (clean) speech.

In [14], Wu et al describe small but important modifications to Han's system, namely substituting a linear activation function for the output layer of the DNN and normalizing the target features to zero mean and unit variance, but also using a larger number of spectral magnitude values per frame. They then proceeded to develop a so-called reverberant time aware (RTA) DNN which processes input

Paper Approved for Public Release; Distribution Unlimited: 88ABW-2017-3103 20170623

http://dx.doi.org/10.21437/Interspeech.2017-461

speech according to a pre-estimated T60 value (a measurement of the amount of reverberation time) using a carefully chosen frame shift interval (between 2ms and 8ms) and frame context (up to 5 forward and backward adjacent frames). In our experiments, we have also observed the improvement at different T60 values depending on the context window and framing. This adaptive strategy is treated as independent, as it can also be applied to our proposed method.

2.2. Performance of DNN dereverberation

In the aforementioned works, the authors show that DNN dereverberation processing improves speech quality. The authors did this by training networks to map reverberant spectrograms to clean ones by providing only examples of reverberant speech for training. This appears to be a good strategy to allow the network to compensate for different levels of reverberation, but tests revealed a substantial degradation of speech which was already clean or with very little reverberation, a problem that plagues many speech enhancement methods. This is a particularly important problem for SID tasks. Wu et al partially addressed this by showing that for lower T60 values, their improved network would not degrade the speech as Han's did. However, nearly clean speech is still degraded. This served as motivation to search for a solution to this shortcoming.

Performance of the DNN dereverberation, in this paper, is first quantified using speech quality scores. Specifically, we discuss scores from the objective metrics perceptual evaluation of speech quality (PESQ) [17] and short-time objective intelligibility (STOI) [18]. The primary goal of our work, however, is improving performance on SID tasks, so the emphasis is to show a decrease of the equal error rate (EER) for speaker verification over existing methods.

3. Improving DNN dereverberation

In this work, we sought to improve the performance of SID tasks under reverberant conditions. As with any technique using DNNs, this procedure involves many important details which have great effect. We now propose some detailed modifications to the previous works that were critical to achieving this good performance.

3.1. Important details for improved performance

It is important for successful training that the input and target utterances be well aligned in time to avoid having the DNN be responsible for the added chore of compensating for poor alignment. After convolution with the RIR to add reverberation, the resulting reverberant speech waveform is longer and shifted in time with respect to the initial clean speech waveform. The two waveforms were time aligned by finding the point of maximum correlation between the clean and reverberant sound and shifting the reverberant waveform by this time difference. The tail portion of the reverberant waveform, corresponding to the filter's "ramping down" segment, was then truncated so that the two waveforms were of equal length. This strategy produced considerable improvement over the "theoretically correct" strategy of aligning based on the delay of the direct sound as determined by the location of the first major peak in the RIR.

Another important detail is the amplitude of the speech waveforms, which were scaled so that the maximum values are all equal. Again, this detail boosts performance of the DNN by reducing the need to cope with variability due to speaker volume.

After signal analysis, the input and target spectral features were globally normalized using the well-known mean and variance normalization (MVN) strategy. We also investigated a strategy involving normalization on a finer scale, separate normalization for each utterance or by T60 value to reduce extra variability, but both of these strategies degraded performance significantly and also complicated the process of denormalization processing of the output. So, these methods were not used for any of the results reported in this paper.

When implementing the systems described by Han and Wu for comparison, a sizable increase in performance was observed due to simply adopting a larger fast Fourier transform (FFT) length in the signal analysis step, despite using the same frame size. This suggests that higher spectral resolution is beneficial and we are currently investigating this issue. Larger FFT values improved results, but not surprisingly exhibited diminishing returns, particularly if the size of the network and training set were not taken into consideration. For the work in this paper, a 1024 point FFT was used, corresponding to 513 spectral magnitude features per 32ms Hamming windowed frame, each separated by 16ms spacing.

3.2. Improved training with clean data

Ideally speaking, the DNN should accurately reconstruct the clean speech magnitude spectrum from the reverberant spectrum while also identifying and allowing any clean speech to pass through unaffected. Such a system would greatly improve usability in practical situations. One could achieve a similar effect by pursuing a technique like Wu's RTA-DNN, whereby speech estimated to be clean is simply bypassed without DNN processing. However, the true goal of training such a DNN is to let the network learn about the appropriate correlations between adjacent frames of clean speech so as to correct them when adjacent frames appear too correlated (as is the case in highly reverberant speech). By only training the network with adjacent frames of reverberant speech as input, the network cannot be expected to learn the true distribution of clean frames. For these reasons, it therefore would seem to be useful to include samples of clean speech in the training data.

To that end, we added clean speech segments into the training dataset so that, for those cases, the network presumably learns to perform as an identity function. In this way, the network therefore is exposed to the full gamut of scenarios and better learns the true distribution of adjacent clean features. It also requires no additional data.

3.3. Implementation

In order to make comparisons to the existing work, we attempted to accurately duplicate the data setup first described by Wu in [14], whereby 10 artificial RIRs were generated using the improved image source method (ISM) [19] for a range of T60 values (0.1s to 1s with 0.1s increment) and were convolved with the TIMIT [20] training dataset of 4620 utterances, resulting in about 40 hours of reverberant training speech. The test set for speech quality was a randomly selected set of 100 utterances from the TIMIT test set convolved with 20 generated RIRs with T60s ranging from 0.05s to 1s in 0.05s increments (note the additional test case of 0.05s). Using the entire TIMIT test set for testing produced very similar results. The RIRs were generated according to

Paper Approved for Public Release; Distribution Unlimited: 88ABW-2017-3103 20170623

Wu's description, with a room of size [6x4x3] meters (length by width by height) and speaker and microphone locations of [2, 3, 1.5] and [4, 1, 2] meters, respectively.

The networks were trained with the Microsoft Research (MSR) Cognitive Toolkit (CNTK) [21] to minimize square error between the estimated spectrum and clean spectrum. The network had 3 hidden layers with 2048 nodes in each layer and rectified linear activations. Three forward and backward context frames (7 frames total x 513 values = 3591 total input features) were used. As mentioned above, changing the frame context window increased performance for different T60 times (higher T60 benefits from longer context), but we did not use the T60 adaptive strategy used in Wu's RTA-DNN. The context of 7 total frames was chosen as the best fixed context length. For the initial epoch, we used minibatches of 256 and a learning rate of 0.005 per batch. For the remaining fine-tuning epochs, we used minibatches of 512 with minibatch size adaptation and a learning rate of 0.00005 per batch with a maximum of 20 epochs.

4. Experiments and results

In order to show the effect of the proposed changes, we compared results for 4 different trained networks against other proposed dereverberation methods which also operate as a preprocessing step, namely Temporal Masking and Thresholding (TMT) [6] and Blind Spectral Weighting (BSW) [7]. The first network, which we label "Proposed," implements all items noted in section 3. The second network is identical to the first except the clean training data is absent. The third is our implementation of Wu's first network as described in [14], labelled "Wu et al." Specifically, this does not include T60 adaptive framing (RTA-DNN) because this technique can be applied on top of the proposed modifications. Lastly, we also implemented Han's work of [12] for reverberation. All networks were trained with the described data based on TIMIT. The first network, however, was supplemented with a copy of the clean data. Specifically, the dataset included a total of 50820 utterances (46200 reverberant + 4620 clean). We only plot results from the first and third networks to avoid overcrowding the figures.

Underlying all of this work is the fact that this dereverberation only operates on the magnitude response because humans are relatively insensitive to phase [22]. Therefore the ideal result, which is the best possible result achievable by this processing, is obtained by reconstructing the waveform using the clean magnitude spectrum and the reverberant phase spectrum. We also present this baseline to show the upper limit of possible improvement, though perhaps future work should explore methods for cleaning the phase as well. Note that, as one test of the setup, speech was also synthesized from clean magnitude and phase. As expected, results were the same as for the unprocessed clean speech.

4.1. Speech quality

To illustrate the method, we processed and reconstructed reverberant speech to produce speech quality scores. As an example, Figure 1 shows the reverberant and resultant processed spectrogram for a speech segment from the test set. Note the normally silent gaps between phonemes and the harmonics have been well restored by the DNN. Figure 2 shows the average PESQ scores for the entire test set.

We note an important distinction between Figure 2 and those presented by Wu [14]. Despite replicating their described data setup as exactly as possible, our reference reverberation scores do not match theirs and the scores presented in their work exceed what is theoretically possible (ideal result) for the dataset they describe. Since these curves are based on the RIRs and the data set (TIMIT, which we have matched), the discrepancy seems to lie in the RIRs.

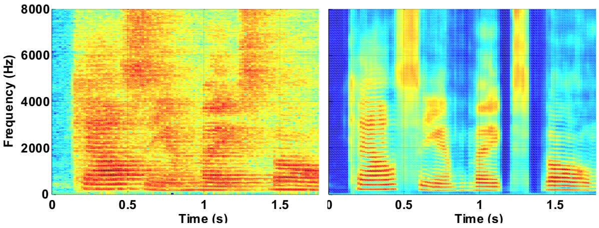

Figure 1*: Example of reverberant speech with T60 of 1s (left) and DNN processed speech (right)*

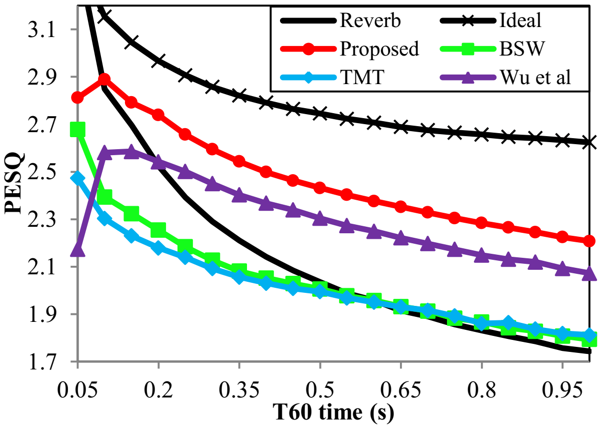

Figure 2*: Average PESQ for Proposed method, our implementation of Wu et al, TMT, BSW, and baselines*

As can be seen in Figure 2, our method improves speech quality over Wu's for all tested T60 values. We especially note the large improvement at very low T60 times (< 0.2s) due to including clean data in the training. Our network trained without clean data also increases scores at all T60 values compared to Wu's, but less so for low T60 values. This is an important point, considering that training with clean data does, in a manner of speaking, increase the size of the training set even though no new data is actually used. Han's network does increase speech quality scores over the unprocessed ("Reverb") speech, but not for smaller T60 values as Wu also noted. The scores from Han's network are lower than the Proposed and Wu's, but still mostly better than BSW and TMT, which can be seen to improve scores only in stronger reverberation scenarios (T60 > 0.55s). We also computed scores for the STOI metric, which showed a similar ranking.

Of the modifications proposed, the largest improvement came from increasing the FFT length (512 to 1024), which increased average PESQ scores for all T60s by about 3% over Wu et al. Correlating and time aligning the waveforms produced another 1% improvement. Scaling the amplitude of the waveforms produced another 2% improvement. Finally, the addition of clean training data increased overall average performance by about 1%, but nearly all this improvement was for T60 values under 0.25s (3.4% gain) versus negligible changes for other T60 values. The average total PESQ improvement over all sentences/conditions was 0.16 and

Paper Approved for Public Release; Distribution Unlimited: 88ABW-2017-3103 20170623

96.4% of test files were improved compared to Wu's method.

4.2. Speaker verification

In order to show the effect of the described DNN dereverberation processing for the task of speaker verification, a series of experiments were performed. Training SID models with data that contains reverberation can help improve performance on test speech that also contains it, but a preprocessing step which can transform degraded speech into clean speech is ideal. Therefore, we trained the SID models using clean data and compared performance on test data processed by the dereverberation methods.

All experiments were done with the TIMIT database. The SID system used was the Alize [23] iVector system with probabilistic linear discriminant analysis (PLDA) scoring. We used a 1024 mixture universal background model, iVector dimension of 200, and a PLDA Eigenvoice and Eigenchannel dimension of 100 and 50, respectively. The configuration we used for speaker verification contains 5300 utterances for training of the background model and scatter statistics. The remaining 1000 utterances in the database were partitioned at 900 and 100 for enrollment and test, respectively. The test set contained 100 speakers (1 sentence per speaker); there were 9 enrollment sentences per speaker. The test and enrollment speakers are not present in the background training set.

The enrollment and test datasets were produced by convolving the clean waveforms with newly generated RIRs (i.e., different than those used for training the DNN). We generated 2 sets of 5 RIRs for T60 values 0.2s to 1s in 0.2s increments where we used randomly generated room sizes and microphone/loudspeaker positions. Each RIR in the first set was used to generate a matched enrollment and test dataset at one T60 value. We label this arrangement our "Matched" experiment. For our so-labelled "Unmatched" experiment, each waveform in the enrollment and test sets was produced by convolving with 1 randomly selected RIR from the 5 RIRs in the second set. We again trained a network for dereverberation in the same manner as described earlier, but this time used the utterances that correspond to the speaker verification training set to ensure no overlap with the test set.

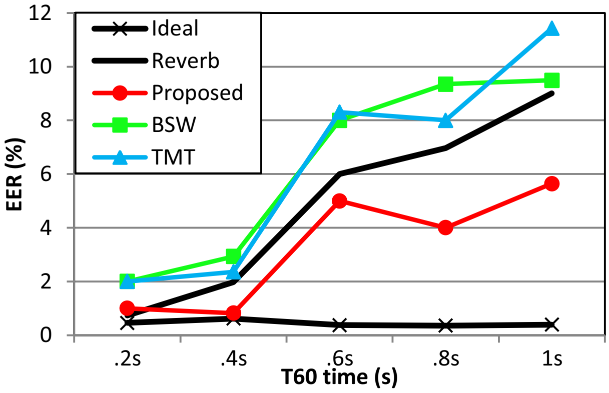

Figure 3*: EERs for Matched experiment*

The enrollment and test data, which was corrupted by varying levels of reverberation, was then processed by the 3 dereverberation methods and features were computed for the speaker verification test. The features used were Mel-Frequency Cepstral Coefficients (MFCCs) [24] 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. Figure 3 and Table 1 depict the EERs for the matched and unmatched experiments, respectively.

These tests show that the proposed method improves verification performance on reverberant speech for a range of T60 values. The error rate for the matched test with a T60 of 0.2s was slightly degraded by 0.26%, which is the only condition tested that did not show significant improvement. The second point to be made is that the proposed method is superior in performance over the other two dereverberation methods. TMT and BSW both degrade performance compared to the baseline for all tested conditions. Note that in other SID work with telephone bandwidth speech, we have observed significant benefit while using TMT and BSW, particularly for highly reverberant conditions and when the algorithms' parameters are tuned to the condition. This particular test therefore shows a significant advantage of the proposed method. It can be used in a more ideal context, without additional tuning, improving performance for varying levels of reverberation when models are trained with clean data. Lastly, note that the features used are based on the magnitude spectrum of the speech, so the "ideal" result is, unsurprisingly, rather unaffected by reverberation. Therefore, future improvements to the method which can better restore the clean magnitude spectrum should result in decreased EERs.

Table 1 : *EERs for Unmatched Experiment*

Ideal	Proposed	Reverb	BSW	TMT
0.5%	6.8%	9.2%	12.7%	11.0%

5. Conclusion

In this paper, a new method for training DNNs for dereverberation was proposed. Important modifications of existing techniques were described which help make a DNN better able to learn how to restore speech contaminated by reverberation. Primarily, preprocessing of the training data to eliminate unnecessary variability allows only the most relevant information to be presented to the network for learning. Additionally, clean data is used in the training dataset to more thoroughly train the network without requiring acquisition of any more actual data. This better allocation of resources improves the training process. Tests were performed to show that these modifications improved the DNN learning and therefore provided improved speech quality.

In this paper and for the first time, this dereverberation method was applied and tested in the context of SID. Preprocessing of speech to optimally improve speech quality may not necessarily lead to optimal performance for other speech processing tasks. Therefore, a set of experiments was performed which showed that the proposed technique is extremely effective at producing improved speech which decreases error rates for the task of speaker verification. Future work will concentrate on improving the method and further investigate some other issues discussed in this paper.

6. Acknowledgements

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. 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.

Paper Approved for Public Release; Distribution Unlimited: 88ABW-2017-3103 20170623

7. References

[1] P. Assmann and A. Summerfield, "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, New York, 2004.

[2] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B. H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Acoustics, Speech and Signal Processing*, 2008.

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

[4] S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver and M. Esnaeilli, "Single-microphone LP residual skewness-based inverse filtering of the room impulse response," in *Audio, Speech, and Language Processing*, 2012.

[5] A. Jukic and S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

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

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

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

[9] I. Peer, B. Rafaely and Y. Zigel, "Reverberation matching for speaker recognition," in *Acoustics, Speech and Signal Processing*, 2008.

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

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

[12] 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.

[13] B. Wu, K. Li, M. L. Yang and C.-H. Lee, "A study on target feature activation and normalization and their impacts on the performance of DNN based speech dereverberation systems," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf*, 2016.

[14] 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.

[15] B. K. A. M. B. R. T. N. Sainath, "Learning filter banks within a deep neural network framework," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.

[16] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier Transform," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.

[17] 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 *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC 2014)*, Antibes, France, 2014.

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

[19] 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.

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

[21] 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.

[22] D. L. Wang and J. S. Lim, "The Unimportance of Phase in Speech Enhancement," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1982.

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

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

Paper Approved for Public Release; Distribution Unlimited: 88ABW-2017-3103 20170623