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

Binary mask estimation strategies for constrained imputation-based speech enhancement

Ricard Marxer[1], *Jon Barker*[1]

[1]University of Sheffield, UK

r.marxer@sheffield.ac.uk, j.p.barker@sheffield.ac.uk

Abstract

In recent years, speech enhancement by analysis-resynthesis has emerged as an alternative to conventional noise filtering approaches. Analysis-resynthesis replaces noisy speech with a signal that has been reconstructed from a clean speech model. It can deliver high-quality signals with no residual noise, but at the expense of losing information from the original signal that is not well-represented by the model. A recent compromise solution, called constrained resynthesis, solves this problem by only resynthesising spectro-temporal regions that are estimated to be masked by noise (conditioned on the evidence in the unmasked regions). In this paper we first extend the approach by: i) introducing multi-condition training and a deep discriminative model for the analysis stage; ii) introducing an improved resynthesis model that captures within-state cross-frequency dependencies. We then extend the previous stationary-noise evaluation by using real domestic audio noise from the CHiME-2 evaluation. We compare various mask estimation strategies while varying the degree of constraint by tuning the threshold for reliable speech detection. PESQ and log-spectral distance measures show that although mask estimation remains a challenge, it is only necessary to estimate a few reliable signal regions in order to achieve performance close to that achieved with an optimal oracle mask.

Index Terms: speech recognition, speech enhancement, imputation

1. Introduction

Speech enhancement is a well-established research topic in the field of robust speech processing [1]. The goal is to process a noise-corrupted speech signal in such a way as to recover the quality and/or intelligibility of the original utterance. This is challenging because the problem is typically under-determined (e.g., a single channel recording of speech with many additive noise sources) and requires a priori knowledge from models of either the noise or the speech source.

The conventional approaches to enhancement are essentially ‘subtractive’, that is, they operate by trying to remove the noise energy to leave something closer to the noise-free speech. Such approaches typically focus on the noise model and include spectral subtraction [2, 3] and Wiener filtering [4]. These approaches can work well in predictable noise conditions but poor noise estimation can leave residual noise, or worse, lead to ‘over-subtraction’ resulting in highly distracting perceptual artefacts known as musical noise. Further, although these subtractive approaches can increase the perceived quality, they are generally ineffective in increasing speech intelligibility [5].

In recent years a new breed of synthesis-driven approaches has emerged. These approaches take the opposite approach. They rely most heavily on a model of the target speaker and, rather than subtract noise, they attempt to reconstruct the original signal from parameters estimated from the noisy signal. By using a generative model of the clean speech, these approaches can potentially produce a signal totally free of additive noise, however, errors in parameter estimation can lead to other forms of distortion. Synthesis-driven approaches include inventory-based systems [6, 7] which reconstruct the signal from a library of clean speech segments (akin to concatenative speech synthesis [8]) and more purely model-driven approaches [9, 10] which reconstruct a signal from a statistical representation (akin to HMM synthesis [11]). This paper focuses on the latter.

In earlier work we presented an analysis-resynthesis enhancement approach which was based around the missing-data model of robust speech recognition [9]. In this approach, an initial spectro-temporal mask estimate allows a speech state sequence to be estimated using missing-data techniques [12]. Then the same sequence is used to drive an HMM synthesis that re-estimates the masked speech regions, while being constrained to match the speech observations in the non-masked regions. This approach was highly successful in stationary noise conditions in which a simple noise estimate could be used to estimate the necessary mask. However, these are also the conditions in which subtractive enhancement can perform well.

In the current paper we extend our approach to work in more general non-stationary conditions. In order to do this we weaken the dependence on noise-based mask estimation. First, the mask is now only used during resynthesis in order to select spectro-temporal (ST) points that require re-estimating. The analysis stage is replaced with a DNN-based acoustic speech model that has been trained in a multi-condition style. Second, we experiment with mask estimation techniques that draw from the speech model by using the estimated speech sequence to determine which observed ST points are noise dominated.

This paper is also related to other recently-published HMM-driven enhancement approaches (e.g. [10, 13]). However, these techniques reconstruct the entire time-frequency speech signal rather than using a masking model. We compare full resynthesis to our masked ‘constrained’ resynthesis to test whether adopting the original ST observations in the unmasked regions enables the enhanced signal to more closely approach the original signal. The remainder of the paper is structured as follows. Section 2 presents an overview of the analysis-resynthesis framework. Section 3 presents our set-up for experimental evaluation in which the system is used to enhance a simple sentence embedded in complex domestic audio. Section 4 presents results and Section 5 discussion and conclusions.

This work was supported by EPSRC Grant EP/M026981/1 (AV-COGHEAR). The authors would also like to gratefully acknowledge the contributions of Dr Ning Ma at the University of Sheffield and Pr Dorothea Kolossa at Ruhr-Universität Bochum for providing the Kaldi recipe to train the noisy speech models.

http://dx.doi.org/10.21437/Interspeech.2017-1257

2. Analysis-resynthesis framework

The speech analysis-resynthesis operates in an auditory spectro-temporal domain, i.e., prior to analysis an auditory representation is formed from the noisy signal (Section 2.1) and after processing the resynthesised representation is converted back into a waveform (Section 2.5). The analysis-resynthesis itself is performed by the system presented in Fig. 1. Briefly, during the analysis stage each frame of auditory spectral features is assigned to a latent speech state by means of a state-of-the-art noise-corrupted speech model. In the resynthesis stage each speech state is mapped to a distribution over the clean speech parameters and their time-derivatives. A binary mask is estimated which partitions the spectro-temporal speech parameters into reliable and unreliable parts. Finally, the mask is used to estimate an optimal sequence of speech parameters for all the frames, conditioned on the reliable data.

2.1. Auditory frontend

An auditory front-end is employed to analyze the input signals $\mathbf{x}$ with a bank of $F = 32$ overlapping Gammatone filters, with centre frequencies uniformly spaced on the equivalent rectangular bandwidth (ERB) scale between 100 Hz and 7 kHz [14]. The output of each Gammatone filter $\tilde{\mathbf{y}}_f$ in the auditory front end is framed into T frames $\tilde{\mathbf{y}}_{t,f}$ with frame size M and shift L. A *Hamming* window is applied to each frame and the instantaneous Hilbert envelope is computed, $\mathbf{u}$. The resulting values are log-compressed to obtain an approximation to the auditory nerve firing rate – the 'ratemap' spectro-temporal representation, $\mathbf{c}$ [15]. Note that the feature sequences for the analysis $\mathbf{c}^a$ and resynthesis $\mathbf{c}$ stages can differ.

2.2. Analysis

During the analysis stage each frame $\mathbf{c}^a_t$ is assigned a state $q^a_t \in Q$ by decoding the most probable word sequence $\mathbf{w}$ using the Viterbi approximation:

$$\begin{aligned}\hat{\mathbf{q}^\mathbf{a}} &= \arg\max_{q\in\Pi} \{p(\mathbf{q}|\hat{\mathbf{w}})p(\mathbf{c}^a|\mathbf{q})\} \\ \hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \{p(\mathbf{w}) \arg\max_{q\in\Pi} \{p(\mathbf{q}|\mathbf{w})p(\mathbf{c}^a|\mathbf{q})\}\}\end{aligned} \tag{1}$$

where Π is the space of possible state sequences, $p(\mathbf{w})$ is the prior on word sequences and $p(\mathbf{c}^a|\mathbf{q})$ is the noise-corrupted speech acoustic model.

The space of possible states Q is built by training a GMM-HMM speech recognition model using triphones as state candidates and clustering states with the same central phone to maximize the likelihood. The acoustic model $p(\mathbf{c}^a|\mathbf{q})$ is implemented with a state-of-the-art DNN as described in Section 3.1.

2.3. Enhancement

The speech enhancement process is performed as described in [9]. Given the sequence of T states $\hat{\mathbf{q}} = [q_1, ..., q_T]$ we estimate the sequence of enhanced ratemap frames $\hat{\mathbf{c}} = [\mathbf{c}_1, ..., \mathbf{c}_T]$ by solving for each utterance:

$$\begin{aligned}\hat{\mathbf{c}} &= \arg\max_{\mathbf{c}} \{p(\mathbf{o}|\hat{\mathbf{q}})\} \\ \text{s. t.} \quad & \mathbf{A}\mathbf{c} = \mathbf{c}^r\end{aligned} \tag{2}$$

$p(\mathbf{o}|\hat{\mathbf{q}}) = \prod_t p(\mathbf{o}_t|\hat{q}_t)$ being the acoustic model of the clean speech under the HMM assumption, with $\mathbf{o} = \mathbf{W}\mathbf{c} = [\mathbf{o}_1, ..., \mathbf{o}_T]$ the speech parameters extended with 1st and 2nd temporal derivatives $\mathbf{o}_t = (\mathbf{c}_t, \Delta^1\mathbf{c}_t, \Delta^2\mathbf{c}_t)$. For a given utterance, $\mathbf{W}$ is constructed from the derivative operations Δ^1 and Δ^2 as detailed in [16]. $\mathbf{A}$ is a segregating matrix that selects the spectro-temporal regions considered reliable.

We assume each state output distribution is a single multivariate Gaussian:

$$p(\mathbf{o}_t|q_i) = \mathcal{N}(\mathbf{o}_t; \mu_i, \mathbf{\Sigma}_i) \tag{3}$$

where μ_i, and $\mathbf{\Sigma}_i$ are the mean vector and covariance matrix of the state i, respectively.

Solving Eq. 2 using a Lagrange multiplier [9] leads to:

$$\begin{aligned}\hat{\mathbf{c}} = &\underbrace{\left(\mathbf{W}^\intercal\mathbf{U}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\intercal\mathbf{U}^{-1}\mathbf{m}}_{(1)} \\ &+ \underbrace{\left(\mathbf{W}^\intercal\mathbf{U}^{-1}\mathbf{W}\right)^{-1}\mathbf{A}^\intercal\eta}_{(2)}\end{aligned} \tag{4}$$

$$\begin{aligned}\eta = &\left(\mathbf{A}\left(\mathbf{W}^\intercal\mathbf{U}^{-1}\mathbf{W}\right)^{-1}\mathbf{A}^\intercal\right)^{-1}\mathbf{c}^r \\ &- \left(\mathbf{A}\left(\mathbf{W}^\intercal\mathbf{U}^{-1}\mathbf{W}\right)^{-1}\mathbf{A}^\intercal\right)^{-1}\mathbf{A} \\ &\times \left(\mathbf{W}^\intercal\mathbf{U}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\intercal\mathbf{U}^{-1}\mathbf{m}\end{aligned} \tag{5}$$

where $\mathbf{m} = [\mu_{q_1}, ..., \mu_{q_T}]$ and $\mathbf{U} = diag[\mathbf{\Sigma}_{q_1}, ..., \mathbf{\Sigma}_{q_T}]$ are mean vector and covariance matrix for state sequence $\mathbf{q}$ [16].

Term (1) in Eq. 4 depends exclusively on $\mathbf{m}$ and $\mathbf{U}$ which are estimated from $\mathbf{q}$ and the clean speech model, this term is also the *resynthesis* solution $\hat{\mathbf{c}}^{\text{R}}$ to the problem if no reliable data is present. The term (2) depends on the segregation matrix $\mathbf{A}$, and is responsible for anchoring the reliable regions $\mathbf{c}^r$ in the constrained imputation solution $\hat{\mathbf{c}}^{\text{CI}}$.

2.4. Mask estimation

The role of the mask is to classify the static speech parameters $\mathbf{c}$ into two classes: those that are considered to be reliable speech observations, $\mathbf{c}^r$ and those that are considered unreliable $\mathbf{c}^u$ (i.e., masked by noise).

In this work we consider two contrasting approaches to mask estimation. First, a mask derived from an estimate of the noise employing an assumption of stationarity. Second, a mask derived from our model of the clean speech. We also generate oracle results by employing an a-priori mask using ground truth information about the clean speech and noise signals.

The masks are computed in two steps. We first produce a time-frequency SNR map by estimating a signal-to-noise ratio (SNR) value $\mathbf{s}_{t,f}$ for each time-frequency cell in the representation (see below). We then produce a binary mask by applying a threshold SNR to the map, i.e., a bin is considered as reliable if its estimated SNR is above the threshold. Instead of setting an SNR threshold explicitly, in our experiments we tune the threshold SNR to achieve a given proportion of reliable bins (r_{ratio}). (Note, this proportion has been called the 'glimpse proportion' in models of speech perception that measure masking in the same auditory domain [17]).

The SNR map of the *noise-based mask* is estimated under the assumption that the noise remains stationary and that we have access to a speech-free segment of noise prior to the start of the utterance. Specifically, an estimate of the noise energy is constructed by averaging the ratemap representation over a 100 ms period ($B = 20$ frames) prior to the speech onset time.

The *speech-based mask* exploits the fact that we have a model of the clean speech. To construct this mask we create an

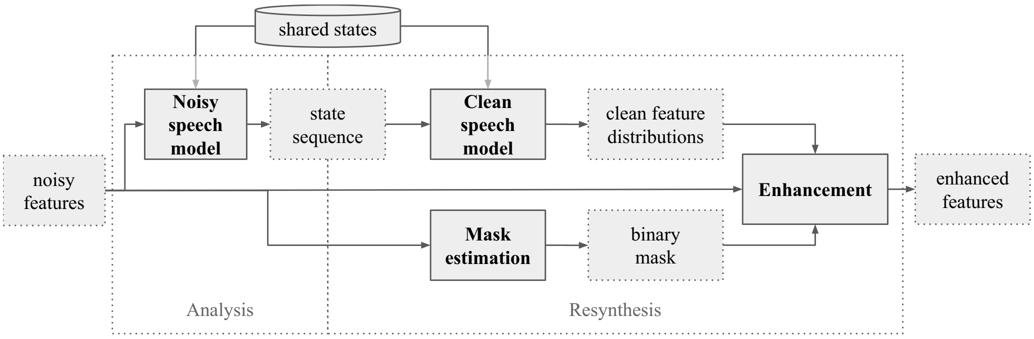

Figure 1: *System overview.*

estimation of the speech without a mask, computing the *resynthesis* solution $\hat{\mathbf{c}}^{\text{R}}$ using term (1) Eq. 4. We then use the speech estimate to derive the SNR map.

The *oracle mask* corresponds to the mask derived from the ground truth SNR map. This is computed from ratemaps of the clean speech and the noise prior to mixing.

The SNR maps of the *noise-based* $\mathbf{s}_t^n$, *speech-based* $\mathbf{s}_t^s$ and *oracle* $\mathbf{s}_t^o$ masks are defined as follows:

$$\hat{\mathbf{u}}_t^n = \frac{1}{B}\sum_{1}^{B}\mathbf{u}_{-i}$$
$$\mathbf{s}_t^n = \frac{\mathbf{u}_t - \hat{\mathbf{u}}_t^n}{\hat{\mathbf{u}}_t^n}, \quad \mathbf{s}_t^s = \frac{\hat{\mathbf{u}}^{\text{R}}}{(\mathbf{u}_t - \hat{\mathbf{u}}^{\text{R}})}, \quad \mathbf{s}_t^o = \frac{\mathbf{u}_t^c}{\mathbf{u}_t^n} \tag{6}$$

where $\mathbf{u}$ are the uncompressed ratemaps, $\mathbf{u}_{-i}$ is the i^{th} frame prior to the speech region of the utterance, $\hat{\mathbf{u}}^{\text{R}}$ is the uncompressed version of the *resynthesis solution* $\hat{\mathbf{c}}^{\text{R}}$ (term (1) of Eq. 4), $\mathbf{u}^c$ and $\mathbf{u}^n$ are the uncompressed ratemaps of the clean and noise-corrupted speech respectively.

2.5. Waveform synthesis

In the current work we focus on the reconstruction of the speech features with different mask estimation strategies. However the audio may be resynthesized from the ratemap features as done in [9] by weighting the windowed output of each Gammatone filter by the ratio between the enhanced and observed uncompressed ratemaps, and applying overlap-and-add on the sum over the filters:

$$\hat{\mathbf{x}}[n] = \sum_{m=0}^{M-1}\sum_{t=1}^{T} h[n - L(t-1) - m]\sum_{f=1}^{F}\frac{\hat{\mathbf{u}}_{t,f}}{\mathbf{u}_{t,f}}\tilde{\mathbf{y}}_{t,f}[m] \tag{7}$$

where the overlapping window $h[m] = 0$ outside of $[0, M-1]$. L is the frame shift,

3. Experimental setup

Experiments were conducted using noisy speech from the 2nd 'CHiME' Challenge (CHiME-2) dataset [18]. This consists of utterances from the Grid corpus [19] mixed into binaurally-recorded domestic noise backgrounds. For the work here we down-mixed to a single channel. The acoustic models are constructed using the full 17,000-utterance training set. The enhancement is evaluated using the development set for which we have pre-mixed noise and reverberant speech signals. Given the focus of this work, the reverberated speech is considered as the clean signal.

Evaluation is performed using the 9 dB, 3 dB and -3 dB CHiME SNRs. Note, in CHiME, SNRs is controlled by mixing the utterances into sections of the background that have the corresponding level, i.e., the -3d dB setting is not only noisier than the 9 dB setting, it also has a very different noise characteristic (less stationary, and more unpredictable events).

To reduce computational costs, the experiments have been conducted on a 10% random subsample of the utterances of each speaker in the development set. This resulted in a set of 366 utterances for testing purposes.

We compute the Mean Square Error (MSE) between the ratemaps of the clean speech signal and those of the noise corrupted audio. To reduce the effect of boundary artefacts and silence preceding and following the speech we omit the first and last 50 ms segments from evaluation. We also evaluate the synthesized waveform against the clean signal using PESQ.

3.1. Noise-corrupted speech model

The ratemaps for the analysis stage $\mathbf{c}^a$ were computed with a frame length of 20 ms and a shift of 10 ms. The acoustic modelling of the analysis stage $p(\mathbf{c}^a|\mathbf{q})$ consists of a DNN with a Time Delay Neural Network (TDNN) architecture using i-vectors as auxiliary features [20]. We use the TDNN architecture and training procedure presented by [21] to estimate the HMM-state posteriors for each frame of the audio stream. In this architecture, each hidden layer takes as input the concatenation of the previous layer's output at multiple time steps. The input layer of the network receives the LDA transformation of the feature stream spliced with two frames at each side. The neural network consists of seven hidden layers. The indices of the time steps concatenated at each hidden layer are: -4 to 4 for the first layer, -2 and 2 for the third and -4 and 4 for the fifth. A p-norm non-linearity is used for neurons activations [22], with p=2, an input dimension of 850 and an output of 170.

The training strategy is similar to that used for mask estimation. Hidden layers are added gradually every two epochs. During each epoch a batch of 512 samples was used. The effective learning rate was gradually decreased from 0.005 to 0.0005. The training consisted of 4 epochs. Training is performed in parallel using natural gradients and parameter averaging as described in [23]. The LDA transformation and the DNN are trained using an alignment produced by the baseline GMM system, where the number of estimated clusters of states resulted in $|Q| = 1453$. The setup and training procedure is implemented in Kaldi [24] and was adapted from the one employed in [25].

The results presented use the Grid grammar as language model $p(\mathbf{w})$ when decoding the utterances at test time. The

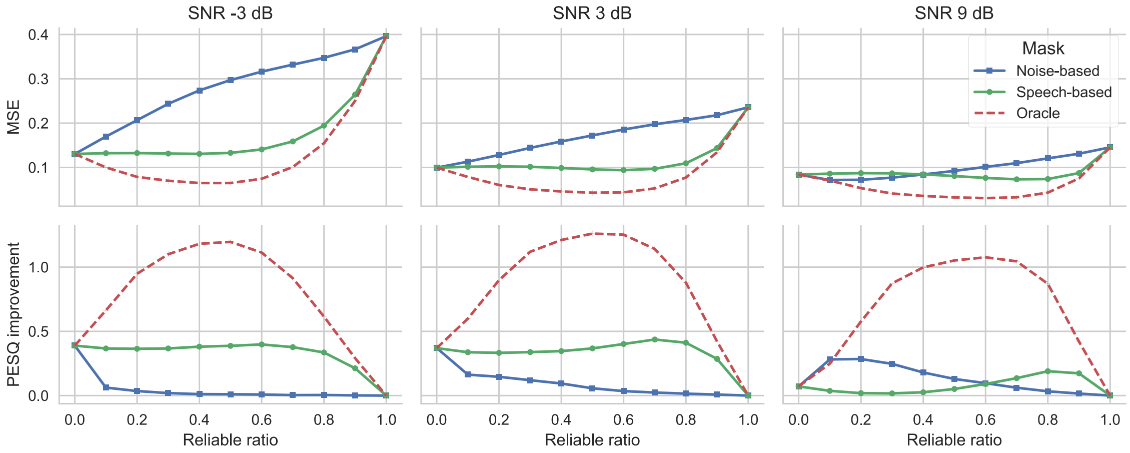

Figure 2: *The Mean Squared Error and PESQ improvement of the enhanced ratemaps for different mask strategies in relation to the ratio of reliable time-frequency bins in the mask.*

results with the ground truth word sequences (not shown here) are very similar. On the development set, the decoder achieves 86.9% in letter and digits word accuracy (the standard performance measure for the Grid corpus – this is comparable with the state-of-the-art on this task (i.e., see [26]).

3.2. Clean speech model

In the resynthesis stage the frame length and shift of the ratemaps $\mathbf{c}_t$ were set to 10 ms and 5 ms respectively. Given the difference in rates between $\mathbf{c}_t^a$ and $\mathbf{c}_t$, the state sequence used in resynthesis is an upsampled version of the analysis state sequence $\mathbf{q} = [q_1^a, q_1^a, ...q_{T^a}^a, q_{T^a}^a]$.

The acoustic model of the clean speech $p(\mathbf{o}_t|q_i)$ for each speech state $q_i \in Q$ is trained on the official CHiME-2 training set using the noisy speech model in order to produce frame-level state alignments. We use the resulting alignments with the ratemaps of the clean version of each utterance to train an individual model for each state. Diagonal covariance matrices $\mathbf{\Sigma}_i$ are used to limit the number of parameters.

We first train a speaker independent model using the utterances from all speakers. Speaker dependent models are constructed using the 500 training utterances of the speaker, falling back to the speaker independent model with states for which less than 20 frames of data are available. In our experimental set-up there is no utterance overlap between training and test sets.

4. Results and discussion

In Fig. 2 we present the MSE values and the PESQ *improvement* of the *constrained imputation* solution $\hat{\mathbf{c}}_{\text{CI}}$ for different SNR values, reliable component ratios (r_{ratio}) and masks (*noise-based*, *speech-based* and *oracle*). Note, a reliable component ratio of 1.0 means performing no enhancement, since all components are considered reliable already, whereas, a reliable component ratio of 0.0 is equivalent to the unconstrained resynthesis solution $\hat{\mathbf{c}}_{\text{R}}$, i.e., comparable to works such as [13].

In all cases the unconstrained resynthesis greatly reduces the MSE with respect to the original signal. In particular, note that the MSE for the processed speech at -3 dB is lower than that of the unprocessed speech at 9 dB, i.e., the enhancement has reduced distortion in the ratemap domain by an equivalent to a 12 dB noise reduction. Note also, however, that for the unconstrained resynthesis distortion remains above an MSE value of about 0.08 even at 9 dB. Although the unconstrained model synthesises a ratemap from clean speech parameters, it will be drawn towards the mean (i.e., typical) rendering of the utterance. This distortion towards the mean will reduce natural variability from the original instance.

Results for constrained resynthesis show that there is potential to reduce MSE further. When using an oracle mask with a reliable ratio of around 0.5 the MSE values are less than half those observed in the unconstrained case. However, trying to meet this performance using estimated masks proves challenging. The poorly estimated noise-based mask generally performs worse than pure resynthesis at all ratios, except at 9 dB where a small gain can be seen with a mask tuned to 0.2. This is explainable by the fact that the 9 dB noise setting in the CHiME data is relatively stationary meeting the assumption of the noise model. However, the speech-based mask is more promising. Although it does not meet the oracle performance, over a wide range of ratios its performance is no worse than pure resynthesis and in the 3 dB and 9 dB conditions it makes small improvements. Optimum performance appears with a ratio of around 0.8, i.e., re-estimating a small number of the most noise affected ST points.

The figure also shows PESQ improvements computed from the reconstructed signals. There results are largely consistent with those of the MSE measures – reductions in MSE translate to improvement in PESQ. What is surprising is that the PESQ improvements for the resynthesised signal are surprisingly small compared to the potential gains that would be seen by constrained resynthesis if the correct mask is used.

5. Conclusions

We introduced an analysis-resynthesis enhancement system with a DNN-based analysis component and a constrained resynthesis employing an estimated S-T mask. Oracle masks show the potential of the technique but good performance in everyday noise requires accurate mask estimation. A novel speech-model based mask estimator demonstrated small improvements in objective measures over unconstrained resynthesis. Listening experiments are now required to validate the results.

6. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] 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.

[3] R. Martin, "Spectral subtraction based on minimum statistics," in *in Proc. Euro. Signal Processing Conf. (EUSIPCO)*, 1994, pp. 1182–1185.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 47–56, 2011.

[6] X. Xiao and R. M. Nickel, "Speech enhancement with inventory style speech resynthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1243–1257, 2010.

[7] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 822–836, 2011.

[8] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 373–376.

[9] J. L. Carmona, J. Barker, A. M. Gómez, and N. Ma, "Speech spectral envelope enhancement by hmm-based analysis/resynthesis," *IEEE Signal Processing Letters*, vol. 20, no. 6, pp. 563–566, June 2013.

[10] H. Veisi and H. Sameti, "Speech enhancement using hidden markov models in mel-frequency domain," *Speech Communication*, vol. 55, no. 2, pp. 205–220, 2013.

[11] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[12] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech communication*, vol. 34, no. 3, pp. 267–285, 2001.

[13] A. Kato and B. Milner, "Using hidden markov models for speech enhancement." in *INTERSPEECH*, 2014, pp. 2695–2699.

[14] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1, pp. 103–138, 1990.

[15] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[16] H. Zen, K. Tokuda, and A. W. Black, "Review: Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2009.04.004

[17] M. Cooke, "A glimpsing model of speech perception in noise," *J Acoust. Soc. Am*, vol. 119, no. 3, pp. 1562–73, 2006.

[18] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines." in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[20] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 225–229.

[21] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of Interspeech*. ISCA, 2015.

[22] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 215–219.

[23] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of dnns with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.

[24] 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," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[25] H. Meutzner, N. Ma, R. Nickel, C. Schymura, and D. Kolossa, "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," in *Proceedings of ICASSP*, 2017.

[26] 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 *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 162–167.