AN OPTIMISATION APPROACH FOR ENHANCING SPEECH INTELLIGIBILITY USING TIME-VARYING SPECTRAL SHAPING IN NOISE

Maryam Al Dabel[1], Jon Barker[2]

[1]College of Computer Science and Engineering, University of Hafr Al Batin, KSA

[2]Department of Computer Science, University of Sheffield, UK
maldabel@uhb.edu.sa, j.p.barker@sheffield.ac.uk

ABSTRACT

We develop an optimisation approach to near-end intelligibility enhancement based on a time-varying spectral shaping to increase speech intelligibility in non-stationary noise. It works by modifying the first few coefficients of an auditory cepstral representation such as to maximise an intelligibility metric using noise-related information. We experiment with two contrasting intelligibility metrics. The first is a glimpse-based objective intelligibility metric that is derived from an energetic masking model. While, the second is a discriminative intelligibility metric building on the principles of missing data speech recognition, to model the likelihood of specific phonetic confusions that may occur when speech is presented in noise. The latter metric is computed using a statistical model of speech from the speaker that is to be enhanced. A formal listening test confirm the efficiency of the proposed algorithm in enhancing speech intelligibility in noise.

Keywords: Speech technology, speech perception, objective intelligibility models, intelligibility enhancement.

1. INTRODUCTION

The field of speech intelligibility enhancement, also referred to as near-end intelligibility enhancement, has gained much recent interest. While conventional speech enhancement techniques (e.g. [11]) process the corrupted speech signal *at the receiver-side* of the communication chain, the intelligibility enhancement techniques (e.g. [1, 16, 17, 18, 20, 21, 24]) pre-process the clean speech *at the transmitter-side* before it is broadcast into the noisy environment. This paper focuses on the latter approach using an analysis-modification-resynthesis framework to promote speech intelligibility in the presence of noise.

There are many existing intelligibility enhancement approaches that apply spectral-only modification algorithms in an attempt to imitate the empirically-observed behaviour of speech produced in noise (i.e. the Lombard speech [15, 23]) where more energy is found at higher frequencies [10, 12, 23]. These algorithms are typically generalisations of high-pass filtering, spectral tilt or centre of gravity changes that work by optimising an objective metric of intelligibility for the purpose of minimising the effect of energetic masking (EM) in a known noise scenario (e.g. [14, 19, 22]).

In our earlier work [1, 2], an algorithm was proposed to optimally modify the spectral shape by adapting the first few coefficients of an auditory cepstral representation such as to maximise an intelligibility metric subject to an energy constraint. Two objective metrics were compared, namely the Glimpse Proportion (GP) [4] and the discriminative microscopic intelligibility (DIS) [1]. The GP is computed as the proportion of the spectro-temporal (S-T) representation of the speech signal that is free from masking. This approach maximises the unmasked acoustic features of speech without paying attention to their relative perceptual importance. In contrast, the DIS approach emphasises important acoustic features that are believed to underlie the intelligibility of speech in noise by focusing on phonemes classes. It employs a statistical speech model and is computed as the ratio of probability between the correct transcription and the most probable incorrect transcription (as a single candidate). It requires a pre-trained speaker-dependent speech model and employs missing data speech recognition theory to handle energetic masking [6].

In this work, we aim to advance the work in [1, 2] by proposing a time-varying spectral shaping algorithm in order to account for the masker's temporal fluctuations. The algorithm operates on a sequence of signal segments. Within each segment, it defines a spectral weighting pattern for each S-T element in order to amplify perceptually important S-T ele-

ments above the level of the masker. We hypothesise that applying a temporally-varying spectral shaping algorithm to the target speech is likely to be most beneficial when the masker itself is also modulated, since it will help define an accurate weight for each S-T element.

The remainder of this paper is organised as follows. First, we will explain the general framework of the proposed method in Section 2. Then, a subjective evaluation and comparison to the reference method are provided in Section 3 and Section 4. Finally, the paper is concluded with a discussion and a conclusion in Section 5 and Section 6, respectively.

2. TIME-VARYING SPECTRAL SHAPING

The spectral modification method, proposed in [1], works by defining a weight for each frequency band which remains fixed across time within each band. In this section, we develop a time-varying spectral modification in which a weight is defined for each individual frequency and time-frame.

The signal is considered as a number of temporal segments each with its own shaping, S, controlled by N cepstral parameters $\mathbf{c} = [c_1 \dots c_n]^T$. Specifically, S is defined as

$$(1) \quad S_c(t,f) = \sum_{n=0}^{N-1} a_t c_n \cos(\frac{\pi}{F}(n+\frac{1}{2})f),$$

where $t = 1,\dots,T$, and T is the total number of frames in the segment, $f = 1,\dots,F$ and F is the number of frequency bands and a_t is a time-varying scaling factor shaped as a triangle function to maximises the shaping in the segment centre and reduce it to zero at the segment boundaries to ensure continuity, i.e.

$$(2) \quad a_t = \begin{cases} 2(t_T - t)/T & \text{if} \quad t < t_c \\ 2(t - t_1)/T & \text{if} \quad t \geq t_c \end{cases}$$

Finally, we obtain the modified log spectrogram $\hat{X}(t,f)$ for the segment by summing the original log spectrogram $X(t,f)$ and the spectro-temporal shaping weights, $S_c(t,f)$,

$$(3) \quad \hat{X}(t,f) = X(t,f) + S_c(t,f)$$

Note, this is equivalent to modifying the first N coefficients of the cepstral representation of X.

After spectrally shaping each spectrogram segment, segments are concatenated, the time-domain signal is resynthesised and a level normalisation is applied across the entire signal to ensure that the energy of the signal remains unchanged. The optimisation process follows that described in our earlier time-invariant shaping approach [1] the only difference being the larger parameter space, i.e., whereas before we have a single shaping vector, we now have one per segment.

In this work, segments have been defined by placing segment boundaries at the dips in the signals temporal energy envelope. Roughly speaking, each segment centre will correspond to a vowel centre. This segmentation is motivated by the fact that the temporal envelope variations are reduced in the low-energy parts of speech signal (e.g. nasals, onsets and offsets) [3]. Thus, the shaping will have least impact on the most sensitive parts of the signal, reducing the potential for perceptual distortions.

The parameters for the modification, i.e., the vectors $\mathbf{c}$ which control the shaping of the segments, are then found by optimisation with respect to either the GP or DIS intelligibility measures. Parameter optimisation is performed using the Nelder-Mead Direct Search method [13]. For details see [1].

3. LISTENING TESTS

Twenty four normal-hearing subjects participated in the study. Listeners were students and staff at the University of Sheffield whose age ranged from 18 to 30 years. The listeners were required to be native English speakers, with no history of speech and/or language dis-orders. All were paid for their participation. Ethics permission was obtained following the University of Sheffield Ethics Procedure.

The speech materials are from the Grid corpus [5]. The corpus consists of sentences recorded by a total of 34 native English speakers (18 male and 16 female). All sentences exhibit the same six words with a fixed grammar of the form "*command*" "*colour*" "*preposition*" "*letter*" "*number*" "*adverb*". There are 1000 utterances recorded from each speaker sampled at 25 kHz. The length of each utterance is about 2.2 seconds.

As additive disturbances, two different noises were considered: (i) a stationary speech-shaped noise (SSN) that was generated by filtering white Gaussian noise through a 100-order all-pole filter, the long-term average spectrum of this noise was approximated to match that of the Grid speech material, and (ii) a non-stationary N-talker babble modulated noise (BMN), which were produced by modulating SSN with the envelope of N-talker babble for various N. As in [4], the envelope was calculated by convolving the absolute value of an N-talker babble signal with a 7.2 ms rectangular window. Babble was generated then by summing utterances with equal *rms* energy from the Grid corpus. In this study,

N was set to 5.

Three different speech enhancements are compared in the evaluation, namely; 'TVGP-DRC', 'TVDIS-DRC', and 'SSDRC', and one unmodified natural speech 'ORG'. The 'TVGP' and 'TVDIS' denote the time-varying spectral modification by optimising the GP metric, and the DIS model, respectively. Both are noise-dependent algorithms. We further processed the TVGP and TVDIS modified speech with the time-domain dynamic range compression (DRC) method, as described in [24]. The DRC works by producing a time-varying gain to reduce the envelope variations of a signal. To compare our noise-aware approaches with a state-of-the-art noise-independent approach, we pre-process the clean Grid data with the spectral shaping and dynamic range compression algorithm as described in [24]. We refer to this system as 'SSDRC'

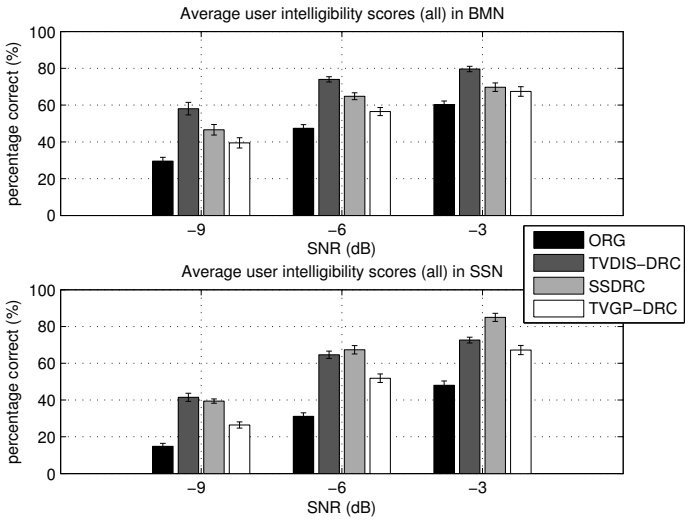
The TVGP and TVDIS were generated with spectral shaping using four cepstral coefficients $\mathbf{c}$ (i.e., $N = 4$). Speech spectral shaping was performed using the filterbank analysis-synthesis framework described in [9]. Note, the first cepstral coefficient, c_0 is arbitrarily fixed to 0 because it simply adds a constant gain factor across frequency that does not change the spectral shape. After resynthesis, the energy of pre-enhanced signal is scaled such that the global signal energy remains unchanged before and after spectral modification. The result is the modified signal, $\hat{x}$, that will be transmitted into the noisy environment.

The SSN and BMN maskers were added separately to the four speech types at three SNRs: -3, -6 and -9 dB. The target utterances were mixed with the masker after the modification mechanism and energy renormalisation.

To train acoustic speech models, a 17,000 utterance training set was provided containing 500 utterances of each of the 34 Grid speakers. We construct phoneme-level HMMs. The number of acoustic phoneme models K is 39. Each phone is modelled using a 3-state HMM with each state modelled as an 7-component diagonal covariance GMM. We first train a speaker-independent (SI) model from the full 17,000 utterances training set. Then we derived a speaker-dependent (SD) model for each of the 34 speakers by running further parameter re-estimations using just the target talker training data.

The four speech types namely: ORG, TVGP-DRC, TVDIS-DRC, and SSDRC, were tested in 3 SNRs conditions of the 2 maskers using a total of $19,584$ stimuli (4 speech types $\times$ 816 utterances (34 speakers $\times$ 24 utterances) $\times$ 6 noise conditions (2 maskers $\times$ 3 SNRs)) divided into independent

Figure 1: The average percentage of utterances in which the letters and digits were identified correctly across listeners as a function of SNR across maskers.



blocks of 136. The independent block was drawn at random, without replacement in which a single subject would hear 34 utterances from each speech types into 6 blocks. The subjects were assigned into blocks in which:

1. each subject heard one block of 136 (34 utterances $\times$ 4 speech types) utterances in each of the 6 noise conditions;
2. no subject heard the same utterance twice;
3. each noise condition was heard by the same number of subject.

Subjects were tested individually in an acoustically-isolated booth. Stimuli were presented once only. The task was to identify the letter and digit spoken and type the heard keywords. Once a participant had typed a response, the subsequent stimulus was presented automatically. Null responses were not permitted. The test was completed on average in 50-60 minutes.

4. RESULTS

In general, performance across listeners was reasonably consistent, so only the mean of the actual identification rates with standard errors averaged across listeners as a function of SNR in the two maskers and the four speech types are plotted in Figure 1. It is evident that, for all modifications, the intelligibility of modified speech signals was substantially higher than that of the original corrupted speech (ORG) across all SNRs for both maskers.

A two-way repeated measure ANOVA with two within-subjects factors (modification type and

Table 1: p-values for comparing intelligibility scores between systems across maskers.

Masker	Methods	ORG	TVDIS-DRC	SSDRC	TVGP-DRC
BMN	ORG	-	0.111	0.145	0.020
	TVDIS-DRC	0.111	-	0.033	0.091
	SSDRC	0.145	0.033	-	0.124
	TVGP-DRC	0.020	0.091	0.124	-
SSN	ORG	-	0.182	0.089	0.098
	TVDIS-DRC	0.182	-	0.092	0.084
	SSDRC	0.089	0.092	-	0.009
	TVGP-DRC	0.098	0.084	0.009	-

masker type) on identification rates indicated statistically significant effects of modification type ($F(3,15) = 21.5,\ p < 0.001$), and the SNR level across maskers ($F(5,15) = 22.7,\ p < 0.001$) on the actual intelligibility of speech in noise. This suggests that the effect of modification strategies varied across SNR and were significantly different from ORG for each masker type.

A post hoc test according to Fisher's LSD ($\alpha = 0.05$), computed separately for each masker type across the SNR level using ANOVAs with the single factor of modification type, indicated several significant differences between the different experimental conditions. The p-values can be found in Table 1. Some p-values appear to be non-significant (e.g. 0.182) due to a very high variance.

The TVDIS-DRC method outperformed all other speech types across SNRs level in BMN. Additionally, the intelligibility gains were roughly the same for the TVGP-DRC and SSDRC. In the SSN masker condition, however, the TVDIS-DRC and SSDRC had a similar pattern of increase at -9 and -6 dB SNR, but the SSDRC (85%) outperformed the TVDIS-DRC (72%) at higher level of SNR. The lowest intelligibility gains were obtained by TVGP-DRC in both maskers and was noticeable in the SSN masker at all level of SNRs.

5. DISCUSSION

All speech types were preprocessed by different spectral modification methods and then by the same time-domain modification, namely the DRC method. In our evaluation, in line with the large scale evaluation in [7, 8], we found that combing the time-domain modification with the time-varying spectral modification resulted in higher intelligibility gain. We attribute the higher gain of intelligibility obtained by TVDIS-DRC to using a better intelligibility-optimisation method that optimised a phoneme-level discriminative microscopic intelligibility. These findings suggested that a significant gain can be achieved by first defining better objective intelligibility metric, and second by combing time-domain modification method.

One limitation of the system is the lack of consistency in the findings across the SNRs which might be associated to the optimisation algorithm. Although the the Nelder-Mead algorithm is appropriate for finding a better solution for the unconstrained problem, it estimates a local maxima based on the current estimates of the simplex. The size and position of the simplex is changing within each alteration of the optimisation which might not alway guarantee the optimal local maxima.

6. CONCLUSION

In this paper, we set out to develop an optimisation approach to near-end intelligibility enhancement which works by exploiting a priori knowledge of a speaker and the noise environment to increase the intelligibility of speech in noise. We automatically modified the speech signal according to the environmental noise by maximising the intelligibility estimate without changing the energy level of speech. We proposed a time-varying spectral shaping, and performed the optimisation on a segment-by-segment basis. Results showed that combining this system with a time-domain noise independent method (i.e. dynamic range compression) improved intelligibility particularly in non-stationary noise compared to the state-of-the-art system.

7. ACKNOWLEDGEMENTS

This research has been sponsored by the Saudi Arabian Ministry of Education and partly supported by the European Community 7th Framework Programme Marie Curie ITN INSPIRE (Investigating Speech Processing in Realistic Environments).

8. REFERENCES

[1] Al Dabel, M., Barker, J. 2014. Speech pre-enhancement using a discriminative microscopic intelligibility model. *Proc. Interspeech.* Singapore, Singapore. 2068–2072.

[2] Al Dabel, M., Barker, J. 2015. On the role of discriminative intelligibility model for speech intelligibility enhancement. *Proc. ICPHS XVIII.* Glasgow, UK.

[3] Allen, J. 2005. *Articulation and intelligibility*. San Rafael, CA: Morgan & Claypool.

[4] Cooke, M. 2006. A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America.* 119(3), 1562–1573.

[5] Cooke, M., Barker, J., Cunningham, S., Shao, X. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America.* 120(5), 2421–2424.

[6] Cooke, M., Green, P., Josifovski, L., Vizinho, A. 2001. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech communication.* 34(3), 267–285.

[7] Cooke, M., King, S., Garnier, M., Aubanel, V. 2014. The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language.* 28(2), 543–571.

[8] Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., Tang, Y. 2013. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication.* 55(4), 572–585.

[9] Hohmann, V. 2002. Frequency analysis and synthesis using a gammatone filterbank. *Acta Acustica united with Acustica.* 88(3), 433–442.

[10] Junqua, J. 1993. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America.* 93(1), 510–524.

[11] Loizou, P. 2013. *Speech Enhancement: Theory and Practice*. CRC Press. 2 edition.

[12] Lu, Y., Cooke, M. 2008. Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America.* 124(5), 3261–3275.

[13] Nelder, J., Mead, R. 1965. A simplex method for function minimization. *The computer journal* 7(4), 308–313.

[14] Petkov, P., Henter, G., Kleijn, W. 2013. Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise. *IEEE Trans. Audio, Speech, Lang. Processing.* 21(5), 1035–1045.

[15] Picheny, M., Durlach, N., Braida, L. 1985. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language and Hearing Research.* 28(1), 96–103.

[16] Sauert, B., Vary, P. 2006. Near end listening enhancement: Speech intelligibility improvement in noisy environments. *Proc. ICASSP.* 493–496.

[17] Sauert, B., Vary, P. 2011. Near end listening enhancement considering thermal limit of mobile phone loudspeakers. *Proc. Conf. on Elektronische Sprachsignalverarbeitung (ESSV), Aachen, Germany.* 333–340.

[18] Taal, C., Hendriks, R., Heusdens, R. 2012. A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure. *Proc. ICASSP.* Portland Oregon, USA. 4061–4064.

[19] Taal, C., Jensen, J. 2013. SII-based speech preprocessing for intelligibility improvement in noise. *Proc. Interspeech.* 3582–3586.

[20] Taal, C., Jensen, J., Leijon, A. 2013. On optimal linear filtering of speech for near-end listening enhancement. *IEEE Signal Processing Lett.* 20(3), 225–228.

[21] Tang, Y., Cooke, M. 2010. Energy reallocation strategies for speech enhancement in known noise conditions. *Proc. Interspeech.* 1636–1639.

[22] Tang, Y., Cooke, M. 2012. Optimised spectral weightings for noise-dependent speech intelligibility enhancement. *Proc. Interspeech.* Portland Oregon, USA.

[23] Van Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., Stokes, M. 1988. Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America.* 84(3), 917–928.

[24] Zorila, T.-C., Kandia, V., Stylianou, Y. 2012. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. *Proc. Interspeech.* Portland, USA.