ON THE ROLE OF DISCRIMINATIVE INTELLIGIBILITY MODEL FOR SPEECH INTELLIGIBILITY ENHANCEMENT

Maryam Al Dabel, Jon Barker

Department of Computer Science, University of Sheffield, UK
mmaldabel3@sheffield.ac.uk, j.p.barker@sheffield.ac.uk

ABSTRACT

This paper uses listening tests to directly evaluate two speech pre-enhancement algorithms that were published in earlier work. The models were previously evaluated using purely objective measures. The methods under study aim to increase speech intelligibility by applying an adaptive spectral shaping filter to the speech signal. In both algorithms the shape of the filter is adapted so as to maximise the intelligibility predicted by an objective intelligibility model. The first algorithm uses 'glimpse proportion' as the measure of intelligibility (i.e. assuming intelligibility is proportional to the extent of energetic masking). In contrast, the second optimises the score of a statistical 'microscopic' intelligibility model that measures the degree of discrimination between the correct interpretation and competing incorrect interpretations of the utterance. Results show that a significant intelligibility gain is obtained when the simple energetic masking model is employed, whereas the discriminative model currently fails to provide any intelligibility improvement.

Keywords: Speech pre-enhancement, objective intelligibility models, intelligibility enhancement.

1. INTRODUCTION

Speech plays a vital role in modern communication systems, e.g. public address systems. Making speech more intelligible in noise is therefore crucial. Typically, speech enhancement techniques (e.g. [11]) are applied to the received signal and attempt to subtract the background noise from the signal. These techniques often improve speech quality but fail to improve intelligibility [12]. A contrasting strategy available in some applications is to *pre-enhance* the speech signal prior to transmission. These pre-enhancement algorithms attempt to modify the clean speech in ways that will protect it from the effects of the predicted noise. In recent years, large advances have been made in the field of developing and evaluating such pre-enhancement algorithms [5, 6, 7].

Pre-enhancement algorithms use a range of techniques (e.g. [17, 18, 19, 21, 22, 24]) but generally they employ strategies that are comparable to those that talkers themselves adopt to counteract the effects of background noise (e.g., the Lombard effect [15]). The acoustic changes in Lombard speech include increasing the fundamental frequency, increasing the intensity, lengthening vowel duration and reducing spectral tilt to boost the high frequencies [10, 13, 23]. To imitate such an effect many pre-enhancement systems use parameters that are tuned using an estimate of the noise context and with the goal of optimising some objective intelligibility measure (e.g. [18, 19, 22]). The weak link is often the quality of the underlying intelligibility model: a highly parameterised enhancement system cannot be reliably tuned against a poorly fitting intelligibility model.

This paper considers intelligibility optimisation based approaches to pre-enhancement that have been recently presented in [1]. These approaches employ a 'microscopic' intelligibility model that is comprised of an auditory 'glimpsing' front-end and a statistical speech model derived from missing data automatic speech recognition (ASR) [4]. The model is termed microscopic because it is designed to make precise predictions about what words listeners will hear when presented with a given noise and speech mixture. In [1] it is argued that a microscopic model may allow speech pre-enhancement systems to be precisely tuned, and a number of objective measures are presented that suggest that the technique has promise. In this paper we present a listening study that directly assess the performance of the approach.

The pre-enhancement approach in [1] optimises the parameters of an adaptive Spectral Shaping Filter (SSF). The SSF has been used in many similar adaptive pre-enhancement systems due to its potential to provide significant intelligibility gains with low computational complexity [6]. In contrast, [24] demonstrates intelligibility improvements using a *fixed* SSF – this fixed SSF is used as the reference method in this paper.

The approach presented in [1] has similarities to

the work of Petkov et al. [14]. Notably, both employ a statistical speech model and attempt to optimise the probability of the noisy signal being recognised correctly. However, whereas the measure in [14] was computed using statistical speech model derived from conventional ASR [16], the work in [1] employs the glimpsing model of auditory masking and a missing data classification scheme that is designed to handle the fact that some spectral-temporal signal regions are masked by noise [4]. Both methods assume the presence of a statistical speech model and that the speech and noise are known in advance.

The remainder of this paper is organised as follows. First, a brief overview of the spectral modification used in this work is explained in Section 2, followed by an evaluation and comparison with the reference method in Section 3 and 4. Finally, a discussion is provided in Section 5.

2. PRE-ENHANCEMENT USING SPECTRAL SHAPING FILTER

The enhancement system uses a Spectral Shaping Filter (SSF) that can acoustically modify the speech in a manner that resembles changes that talkers make naturally when speaking in the presense of noise. [15]. The filter pre-shapes the spectrum of the input speech signal by adjusting the gain of the band-energies. The gains applied to the filters are encoded in the cepstral domain and represented using the first four cepstral parameters. In this paper, we use an adaptive SSF in which these parameters are optimised so as to maximise the predicted intelligibility of the signal.

In earlier work [1], this concept was presented and the tuning of the SSF parameters was performed using two different intelligibility models. In the first method, the intelligibility was assumed to be proportional to the degree of energetic masking as estimated by the 'Glimpse Proportion' (GP) measure [2]. In short, GP uses knowledge of the pre-mixed speech and noise signals, it computes a time-frequnency filterbank representation of each, and computes the proportion of spectro-temporal elements in which the local SNR would be above a 3 dB threshold. In the second, a more complicated model was introduced using missing data classification [4] known as the Discriminative Intelligibility (DI) model which in addition to the speech and noise signals, also employs a speaker-dependent statistical speech model. The DI model uses hidden Markov models (HMMs) to represent the speech. The HMMs are pre-trained using clean speech of the target speaker. Missing data speech recognition techniques are employed and the probability of the correct utterance and the best scoring incorrect utterances is considered. Intelligibility is given by the difference between these scores. For full computational details see [1].

Additionally, a fixed SSF is implemented following [24]. This fixed SSF uses no knowledge of the noise source and serves as a performance baseline. It is motivated by observations in clear speech (formant enhancement [8]) and the reductions of spectral tilt in Lombard speech [13]. It has been demonstrated that this approach is effective in a large-scale open evaluation of speech modification algorithms [6].

3. SUBJECTIVE EVALUATION

In this section we provide the results of a formal listening test. We contrast a number of speech pre-enhancement algorithms and also compare the results to the state-of-the-art. These algorithms are the original (non-modified) speech (ORG) represented the baseline; the GP-based modified speech [1] (GP-OPT); the DI-based modified speech [1] (DIS-OPT); and finally, as a reference algorithm the spectral shaping-based modified speech [24] (SS).

Twenty normal-hearing subjects whose age ranged from 18 to 30 years participated in the listening tests. The subjects 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.

The subjective evaluation of the algorithms was performed using the Grid corpus [3]. The corpus consists of 34 native English talkers speaking simple 6-word command utterances from a fixed grammar (e.g. 'bin green at k zero now'). The algorithms were tested in 5 noise conditions using a total of $13,600$ stimuli (4 algorithms x 680 sentences (34 speakers x 20 utterances) x 5 conditions) divided into independent blocks of 136. The independent block was drawn at random, without replacement in which a single subject would hear 34 sentences from each entry into the 5 blocks (34 x 5 = 170 sentences in total). The subjects were assigned into blocks in which;

1. each subject heard one block of 136 (34 utterances x 4 algorithms) sentences in each of the 5 noise conditions;
2. no subject heard the same sentence twice;
3. each noise condition was heard by the same number of subjects.

Subjects were tested individually in an acoustically-isolated booth. Stimuli were presented once only.

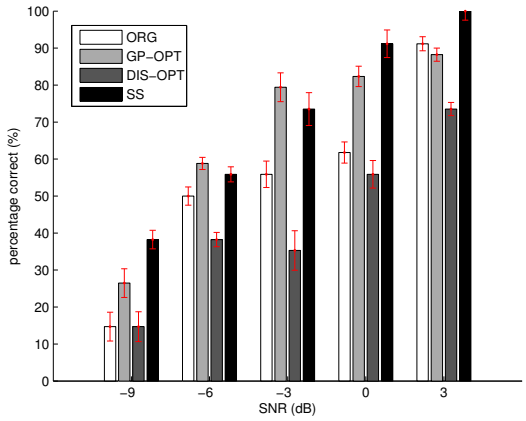

Figure 1: The average percentage of utterances in which the letters and digits were identified correctly across listeners as a function of SNR.

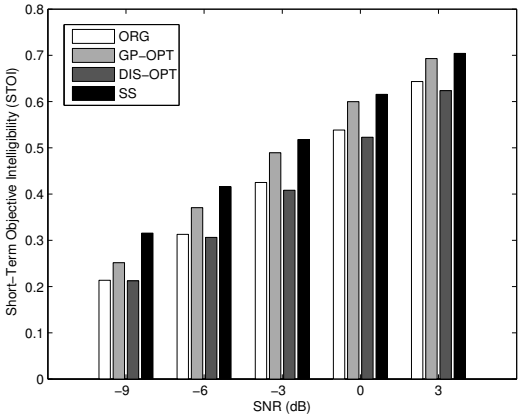

Figure 2: The Short-Term Objective Intelligibility (STOI) results for ORG, GP-OPT, DIS-OPT and SS approaches at a range of SNRs.

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 45 minutes.

The original and modified speech were corrupted by speech-shaped noise masker at a range of SNRs: 3, 0, -3, -6 and -9 dB. This masker was sampled at 25 kHz. The target utterances were mixed with the masker during the testing procedure (i.e. after the modification mechanism) at a desired SNR level.

The speech in DIS-OPT and GP-OPT stimuli was processed using a filterbank Analysis-Modification-Resynthesis framework. First, the speech signal is filtered using a bank of 32 gammatone filters with centre frequencies spread evenly on an equivalent rectangular bandwidth (ERB) scale between 50 and 8000 Hz with filter bandwidths matched to the ERB of human auditory filters. After that, the envelope of each gammatone filter output is computed. This envelope is then smoothed by a first-order low-pass filter with an 8 ms time constant. Then, the smoothed envelope is down-sampled to 100 Hz. Following downsampling, the amplitude envelope is squared and logged to turn the amplitude into the log-energy domain. The spectrum is then shaped by applying a band-dependent scaling to the gammatone filter outputs before re-summing them to form the enhanced signal.

Note, an arbitrary spectral reshaping could be represented as 32 independent scaling factors in the log domain. However, to ensure that the spectral shaping is smooth over frequency we consider only spectral shaping profiles that can be represented using the first N terms of a discrete cosine series, i.e. for 32 bands. In this work N has been set to 4. Further, 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 scaling the filterbank outputs, re-synthesis is employed to generate the spectrally shaped speech signal. Care needs to be taken when summing the bands to compensate for band-dependent phase delays introduced by the analysis. For details see [9]. After re-synthesis the spectrally shaped signal is scaled such that the global signal energy remains unchanged before and after spectral manipulation. The result enhanced signal will be transmitted into the noisy environment.

Some processing artifacts might be introduced by the analysis-resynthesis pathway [9]. In order to control for these, the original unshaped stimuli (ORG), was also passed through the analysis-resynthesis pathway but without any intervening spectral shaping. This allowed us to isolate the intelligibility improvement due solely to spectral shaping.

For comparative purposes, listeners also heard a stimuli processed using the fixed SSF (SS). The fixed filter implementation followed that of Zorila et al. [24].

4. EXPERIMENTAL RESULTS

Figure 1 shows the actual recognition rates together with standard errors averaged across listeners for the four algorithms as a function of SNR. The reported scores were computed as the average of percentage of correctly identified letters and digits. From this data, it is apparent that the performance of GP-OPT and SS is doing similarly well across SNRs.

A two-way repeated measures ANOVA with two within-subjects factors (SNR level and algorithm

Table 1: The p-values for comparing intelligibility rates between techniques across SNRs levels.

	ORG	**GP-OPT**	**DIS-OPT**	**SS**
ORG	-	0.024	0.008	0.020
GP-OPT	0.024	-	0.060	0.015
DIS-OPT	0.008	0.060	-	0.015
SS	0.020	0.015	0.0156	-

type) revealed that a gradual positive impact of SNR level ($F(4,12) = 48.67$, $p < 0.05$), and a small but rather significant effect of algorithm type ($F(3,12) = 16.58$, $p < 0.05$). Our primary purpose is to understand if there is an interaction between these two factors on the overall intelligibility. There was a significant difference between the ORG and the remaining entries. When comparing the performance of ORG against both GP-OPT and SS, for instance, it can be seen that the overall intelligibility rate for higher and lower SNRs levels was a nearly equivalent compared to the ORG with a difference equivalent to about 10 % of performance. However, the performance of DIS-OPT is comparable to the ORG across SNRs except at -3 dB.

A further statistical analysis was carried out using a pairwise comparison analysis. Mean intelligibilities are 54.7 %, 67.0 %, 43.5 % and 71.7 % for ORG, GP-OPT, DIS-OPT and SS across SNRs levels, respectively. A major difference can be seen between GP-OPT and DIS-OPT with difference in mean of 23.5 % and between SS and DIS-OPT of 28.23 %. The p-values can be found in Table 1.

It is interesting to test whether these listening tests results could have been correctly predicted by recent objective measures of intelligibility. Figure 2 shows the results of the Short-Term Objective Intelligibility (STOI) at a range of SNRs. The STOI [20] is a perceptual distortion measure which is designed to evaluate the clean and degraded processed speech. It has been shown a high correlation with speech intelligibility [20]. It can be seen that the STOI measures shown in Figure 2 have a high level of agreement with the actual listening results. STOI correctly predicts that both the SS and GP-OPT techniques will improve intelligibility. However. it predicts an improvement equivalent to a roughly 3 dB increase in SNR, which is a slight underestimate of the improvement observed in the listening data. Further, it predicts the DIS-OPT to have roughly the same intelligibility as the original noisy signal, failing to predict that DIS-OPT will actually decrease intelligibility at -6 and -3 dB SNRs.

5. CONCLUSION AND DISCUSSIONS

In this paper, we have evaluated a number of different SSF systems, all of which aim to increase the intelligibility of the keywords of a target utterance in the presence of speech-shaped noise. In particular, we compare the performance of both fixed and adaptive SSFs and see whether using a priori knowledge of sound sources in the mixture and/or pre-trained speech models results in a larger intelligibility gain.

Our analysis shows that the adaptive SSF using a simple measure of energetic masking is able to improve speaker intelligibility in listening task and obtained a significant improvement over the baseline performance across all SNR conditions. However, the adaptive SSF using a more complicated measure of intelligibility based on a statistical speech model failed to improve intelligibility. The most striking result to emerge is that, in contrast to results reported elsewhere [6], in our experiments the adaptive SSF performed no better than the fixed SSF.

The DIS-OPT system is primarily based on a measure of decoding using the DI model. The DI accounts for the entire target utterance, and is not specific to parts of the utterance to be enhanced although the motivation behind this measure was to discriminate between the correct class with the class that is most acoustically similar (e.g. 'm' versus 'n' or 'b' versus 'v'). The nature of listening task, however, was to identify the letters and digits in the spoken utterance in noise. Hence, the development of the system and the listening task might be not compatible and that might explain the poor performance of the enhancer despite the extra-embedded knowledge.

For the adaptive SSF to operate well in non-stationary noise conditions, one could extend the algorithm by adding a time-varying modification. Additionally, one could replace the spectral modification used in this work with different motivated spectral manipulation e.g. modulation filtering.

6. 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).

7. REFERENCES

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

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

[3] 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, 2421.

[4] 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.

[5] 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.

[6] Cooke, M., Mayo, C., Valentini-Botinhao, C. 2013. Intelligibility enhancing speech modifications: the hurricane challenge. *Proc. Interspeech, Lyon, France*.

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

[8] Hazan, V., Baker, R. 2011. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of Americ* 130, 2139–2152.

[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, 510.

[11] Loizou, P. 2007. *Speech enhancement: theory and practice* volume 30. CRC.

[12] Loizou, P., Kim, G. 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *Audio, Speech, and Language Processing* 19(1), 47–56.

[13] 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, 3261.

[14] Petkov, P. N., Henter, G. E., Kleijn, W. B. 2013. Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise. *Audio, Speech, and Language Processing, IEEE Transactions on* 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.

[16] Rabiner, L. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2), 257–286.

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

[18] 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.

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

[20] Taal, C., Hendriks, R., Heusdens, R., Jensen, J. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *Audio, Speech, and Language Processing, IEEE Transactions on* 19(7), 2125–2136.

[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, 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, 917–928.

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