Speech pre-enhancement using a discriminative microscopic intelligibility model

Maryam Al Dabel, Jon Barker

Department of Computer Science, University of Sheffield, UK

m.aldabel@dcs.shef.ac.uk, j.barker@dcs.shef.ac.uk

Abstract

We propose a new approach for optimally pre-enhancing speech signals for given noise conditions. Like others, we optimise the predicted intelligibility of the signal, however, we employ a statistical 'microscopic' intelligibility model that encodes information about which spectro-temporal speech regions are most informative. Uniquely, our optimisation strategy aims to maximise the discrimination between the correct interpretation and competing incorrect interpretations of the utterance. We present results from studies that use speech-shaped stationary noise maskers and show the new strategy leads to solutions that are more varied than the simple high frequency emphasis employed in many pre-enhancement systems.

Index Terms: Speech pre-enhancement, spectral shaping, objective intelligibility models, discriminative microscopic intelligibility model.

10.21437/Interspeech.2014-470

1. Introduction

When speaking in noisy environments talkers unconsciously monitor their acoustic environment and adjust their speech in ways that maintain its intelligibility (i.e. Lombard speech [1]). In contrast, applications that mix speech with other sound sources or transmit recorded speech into noisy environments do not generally modify it appropriately and hence intelligibility is compromised. Recently, techniques have emerged that address this problem by applying pre-enhancements to speech signals prior to transmission (e.g. [2, 3, 4, 5, 6, 7, 8]). These techniques have been evaluated in [9, 10].

Pre-enhancement generally works by making the speech harder to mask. In Lombard speech this equates to increasing the intensity, increasing the fundamental frequency, lengthening vowel duration and reducing spectral tilt to boost the high frequencies [11, 12, 13]. Pre-enhancement techniques make similar changes although generally a fixed-intensity constraint is applied because it is undesirable to maintain intelligibility by simply boosting the signal energy [14]. The challenge for pre-enhancement systems is to optimally adjust the parameters of the speech modification algorithm. Typically the parameters may be tuned by using knowledge of the background noise and an objective intelligibility model (OIM), i.e. they are adjusted so as to maximise the predicted intelligibility (e.g. [3, 4, 6]). This automated approach allows the adoption of highly flexible enhancement algorithms that can finely control the acoustic characteristics of the speech. However, the parameters can only be tuned to within the precision of the intelligibility model.

Recent advances in intelligibility modelling have introduced approaches based on statistical speech models similar to those used in automatic speech recognition (ASR) [15]. These so-called 'microscopic' intelligibility models allow speaker-dependent models to be built which can accurately predict the impact of specific masking patterns on the intelligibility of specific utterances. In this work we investigate whether these models can allow speech pre-enhancement systems to be more precisely tuned than when using traditional OIMs.

The paper focuses on the situation where the speech signal and the noise background are known in advance, e.g. imagine an audio engineer adding a speech track into the audio mix for a movie scene. Our pre-enhancement approach is based on a parameterised spectral shaping that is applied to the speech signal and which effectively redistributes energy between frequency bands, i.e. unmasking some speech components at the expense of others. The parameters of this enhancement are optimised with respect to two alternative intelligibility models: first, as a baseline, we use a commonly-employed speaker-independent model that is based on a measure of energetic masking (the 'Glimpse Proportion' (GP)) and, second, we introduce a more sophisticated microscopic model trained on speech of the speaker to be enhanced. This latter model is able to select masking patterns that maximise the discrimination between the correct utterance and potential mishearings.

The paper is organised as follows. Section 2 describes the essentials of the baseline pre-enhancement system using the GP intelligibility model. Section 3 explains the novel speaker-dependent discriminative pre-enhancement system. Sections 4 describe experiment and results. Section 5 concludes the paper.

2. Baseline pre-enhancement system

The baseline pre-enhancement system uses an approach similar to that of previously reported systems (e.g. [6]). The speech is passed through a parameterised spectral-shaping system. It is assumed that the speech and noise signals are known and the parameters of the spectral-shaper are optimised to minimise the perceptual masking of the speech signal. The spectral shaping and optimisation stages are described in detail below.

2.1. Spectral Shaping

The spectral shaping is implemented using a filterbank analysis-modification-resynthesis technique. First, the speech signal x 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.

The instantaneous Hilbert 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. After that, the smoothed envelope is down-sampled to 100 Hz. After downsampling, the amplitude envelope is squared and logged to convert the amplitude into the log-energy domain. This stage

yields the Spectro-Temporal (S-T) representation, $\mathcal{X}(t, f)$ that will be used as the basis of the intelligibility models described later.

The spectrum is then shaped by applying a band-dependent scaling to the gammatone filter outputs, x_f, before resumming them to form the enhanced signal. An arbitrary reshaping could be represented as 32 independent scaling factors, w_f, where f represents the frequency channel index, i.e. in the log domain,

$$\log(x'_f(t)) = \log(x_f(t)) + \log(w_f) \tag{1}$$

However, in order 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 channels

$$w_f = \sum_{n=0}^{N} c_n \cos(\pi/32(n + 1/2)f) \tag{2}$$

Note, using this formulation the 32 spectral shaping weights are controlled by N parameters, $c_0, \ldots, c_N$. 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, resynthesis 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 [16]. After resynthesis the spectrally shaped signal is scaled such that the global signal energy remains unchanged before and after spectral modification. The result is the enhanced signal, $\mathbf{x}_c$, that will be transmitted into the noisy environment.

2.2. Optimising Intelligibility

In the baseline system intelligibility is estimated using 'Glimpse Proportion' (GP) [17]. The GP measures the proportion of the S-T representation of the speech signal that is free from masking. Specifically, it is computed as the proportion of S-T elements that have a local dB SNR higher than a given threshold θ. It has been shown to be highly correlated with intelligibility. For a given enhanced utterance, $\mathbf{x}_c$, and noise signal, $\mathbf{n}$, the GP can be written as,

$$GP(\mathbf{x}_c, \mathbf{y}_c) = \frac{100}{TF} \sum_{t=0}^{T} \sum_{f=0}^{F} H\left(\mathcal{X}_c(t, f) - \mathcal{Y}_c(t, f) + \theta\right) \tag{3}$$

in which T and F are the numbers of time frames and frequency bands. The $\mathcal{X}_c$ and $\mathcal{Y}_c$ represent the S-T representation of the enhanced speech at time frame t and frequency band f before and after noise has been added. $H\left(.\right)$ is the Heaviside step function.

The enhancement system can now be optimised by searching for the shaping parameters $\mathbf{c}$ that maximise the GP measure for a given utterance and noise, i.e. the optimal parameter values are given by,

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} GP(\mathbf{x}_c, \mathbf{y}_c) \tag{4}$$

We perform the optimisation using the Nelder-Mead Direct Search method for optimisation [18].

3. Discriminative microscopic intelligibility

The novel enhancement technique optimises the signal with respect to a discriminative 'microscopic' model of speech intelligibility that uses hidden Markov models (HMMs) to represent the speech. The HMMs are trained off-line using clean speech of the target talker. Using missing data ASR techniques [15], for a given noise we can compute the probability for the target utterance given the enhanced signal. The enhancement is then performed by finding the spectral shaping that maximises this probability, i.e. maximising the probability that the model 'hears' the word correctly. Details are given below.

3.1. Theoretical development

The intelligibility model can be explained by considering the task of estimating the phoneme class, q^*, corresponding to a single frame of the observed S-T representation of the noise-corrupted enhanced speech signal, $\mathcal{Y}_c$. This noisy spectrum can be approximated as the element-wise maximum of the enhanced speech spectrum, $\mathcal{X}_c$ and the noise spectrum, $\mathcal{N}$, i.e.

$$\mathcal{Y}_c = \max(\mathcal{X}_c, \mathcal{N}) \tag{5}$$

It is also assumed that the listener is able to make a perceptual segregation of the S-T representation and can estimate which channels are dominated by the noise and which by the speech, i.e. that a foreground/background mask, $\mathbf{m}$, is known. The intelligibility can be modelled as the probability that the phoneme is recognised correctly, i.e.,

$$I_c = p(q^*|\mathcal{Y}_c, \mathbf{m}) \tag{6}$$

The task is now to find the spectral shaping parameters, $\mathbf{c}$, that maximise this intelligibility

$$\hat{\mathbf{c}} = \arg\max_{c} I_c = \arg\max_{\mathbf{c}} p(q^*|\mathcal{Y}_c, \mathbf{m}) \tag{7}$$

I_c is evaluated using a missing data framework. The development follows that of other missing data work (e.g. [19]) but it is repeated here because the difference between speech recognition (finding optimum q) and intelligibility enhancement (finding optimum $\mathbf{x}_c$) introduces an important subtlety.

In the missing data framework the representation of the true signal, here $\mathcal{X}_c$, is treated as a latent variable that we integrate over, i.e.

$$\begin{aligned} I_c &= p(q^*|\mathcal{Y}_c, \mathbf{m}) && (8)\\ &= \int p(q^*, \mathcal{X}_c|\mathcal{Y}_c, \mathbf{m})\, \mathrm{d}\mathcal{X}_c && (9)\\ &= \int p(\mathcal{X}_c|q^*) \frac{p(\mathcal{X}_c|\mathcal{Y}_c, \mathbf{m})}{p(\mathcal{X}_c)}\, \mathrm{d}\mathcal{X}_c && (10) \end{aligned}$$

The mask $\mathbf{m}$ defines a discrete segmentation of the spectrum $\mathcal{X}_c$ into 'present' channels $\mathcal{X}_c^p$ (i.e. those where the speech is not masked and is hence directly observed) and missing channels $\mathcal{X}_c^m$ where the speech is masked but known to have energy less than the observed noisy value $\mathcal{Y}_c^m$. The above equation can then be shown to be,

$$I_c = C \int_{\mathcal{X}_c^m = -\infty}^{\mathcal{X}_c^m = \mathcal{Y}_c^m} p(\mathcal{X}_c^p, \mathcal{X}_c^m|q^*)\, \mathrm{d}\mathcal{X}_c^m \tag{11}$$

The constant C depends only on the observed spectrum, $\mathcal{Y}_c$, the distribution $p(\mathcal{X}_c)$ and the masking pattern $\mathbf{m}$. It can be ignored in missing data ASR because it is the same for all hypothesised states q. However, in our case it cannot be ignored because the observed spectrum, $\mathcal{Y}_c$ is dependent on the parameters c over which we are optimising (i.e. C will not be the same for all evaluations). However, C is hard to evaluate accurately

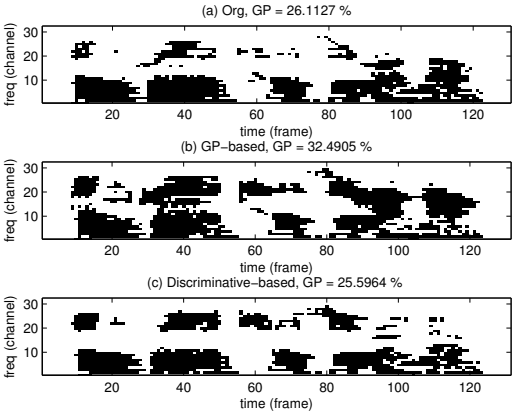

Figure 1: An illustration of glimpse mask representation mixed with artificially generated pink noise at 0 dB for *male* speaker (one utterance): (a) original speech; (b) modified speech using GP-optimised approach; and (c) modified speech using discriminative-optimised approach.

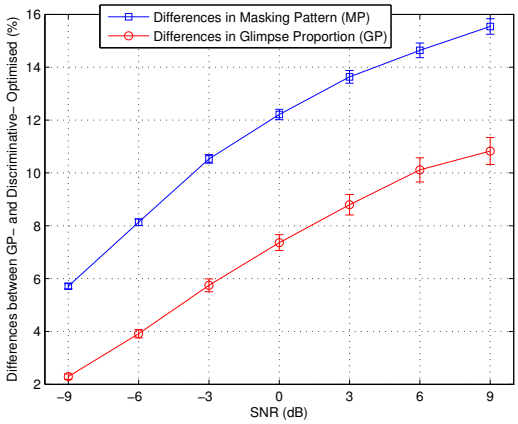

Figure 2: The average differences in masking pattern (MP) and Glimpse Proportion (GP) between GP-optimised and discriminative-optimised approaches at a range of Signal-to-Noise-Ratio (SNR) computed over the full 600 test utterances.

because it requires a good model of the clean speech spectrum prior, $p(\mathcal{X}_c)$.

To avoid needing to evaluate the constant C we can consider a slightly different optimisation problem. Rather than maximising the probability of the correct phoneme we can consider maximising the ratio of probability between the correct phoneme, q^*, and the most probable incorrect phoneme, q^-. This is well motivated because most listening errors will occur not due to accumulated probability of all the incorrect classes but just due to confusing the correct class with the class that is most acoustically similar (e.g. 'b' versus 'v' or 'm' versus 'n').

The intelligibility score that we now optimise is a ratio in which the constant C can be cancelled from the numerator and denominator,

$$I'_c = \frac{p(q^*|\mathcal{Y}_c, \mathbf{m})}{p(q^-|\mathcal{Y}_c, \mathbf{m})} = \frac{\int_{\mathcal{X}_c^m=-\infty}^{\mathcal{X}_c^m=\mathcal{Y}_c^m} p(\mathcal{X}_c^p, \mathcal{X}_c^m|q^*)\,\mathrm{d}\mathcal{X}_c^m}{\int_{\mathcal{X}_c^m=-\infty}^{\mathcal{X}_c^m=\mathcal{Y}_c^m} p(\mathcal{X}_c^p, \mathcal{X}_c^m|q^-)\,\mathrm{d}\mathcal{X}_c^m} \quad (12)$$

3.2. Significant implementation details

The previous section presented the theory for estimating the intelligibility of a phoneme having observed a single spectral feature vector. Clearly, the real problem involves word sequences and a series of spectral feature vectors. The necessary probabilities can be estimated using HMMs to model the sequence data without changing any of the fundamentals of the theory. The states become state sequences: q^*, becomes the most probable state sequence through the known correct word sequence; the competing state, q^-, becomes the most probable sequence through the most probable incorrect word sequence. They can both be estimated using N-best decoding with N set to 2.

The binary mask $\mathbf{m}$ is computed from the known speech and noise signal using the same criterion used to define the S-T glimpses in the previous section, i.e.,

$$m(t,f) = H(\mathcal{X}_c(t,f) - \mathcal{Y}_c(t,f) + \theta) \quad (13)$$

Mask entries equal to 1 indicate the observation is present (i.e. unmasked) and 0 indicates missing (i.e. masked).

It is assumed that spectral shaping does not alter the intelligibility of the clean signal, i.e. it is assumed that listeners can adapt their internal models of speech to fit the shaped signal (this is consistent with common experience). To model this the HMMs are spectrally shaped to match the shaping being applied to the speech. Given that the shaping is a fixed offset to the log spectrum, the models can be adapted by simply adding the same offset to the mean vectors of every Gaussian mixture component of every HMM model state.

Optimisation is performed using the Nelder-Mead Direct Search method [18]. For each iteration the following steps are performed: i/ the speech is spectrally shaped according to the current parameter values, $\mathbf{c}$; ii/ the noise is mixed in; iii/ the noisy representation and mask are computed; iv/ the spectral shaping is applied to the HMMs; v/ the missing data speech recogniser is first run with a grammar fixed to the correct utterance, and then again with the full grammar for the speech corpus being used and with N-best decoding ($N = 2$); vi/ the difference between the log probabilities of the correct and best incorrect decoding is returned as the function value to be optimised.

4. Experimental Results

Experiments were performed using speech data taken from the Grid corpus [20]. The corpus consists of 34 native English speakers (18 male and 16 female) speaking simple 6-word command sentences from a fixed grammar. There are 1000 utterances recorded from each speaker sampled at 25 kHz.

For the discriminative optimisation approach the target speech data is modelled using word-level HMMs. Each word model has 2 states per phoneme (i.e. models vary in length from 2 states to 10 states). The models have a left-to-right no-skip topology. Each state is modelled with a 7 component diagonal-covariance GMM. The training data consist of 500 utterances from each speaker. We first train speaker-independent models using 17,000 utterances (34×500). Speaker-dependent models are then produced by running further parameter reestimations using just the target speaker training data.

The test set contains 600 utterances (roughly 20 per

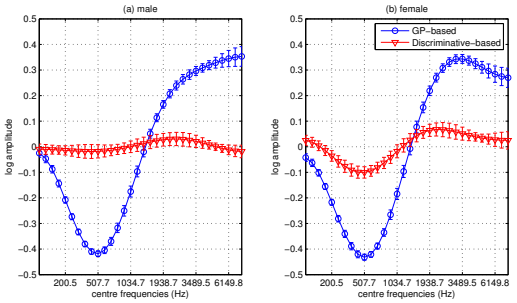

Figure 3: The average optimum spectral shape in spectral domain of the GP-optimised and discriminative-optimised approaches for *male* and *female* speakers separately at 0 dB computed over 100 test utterances.

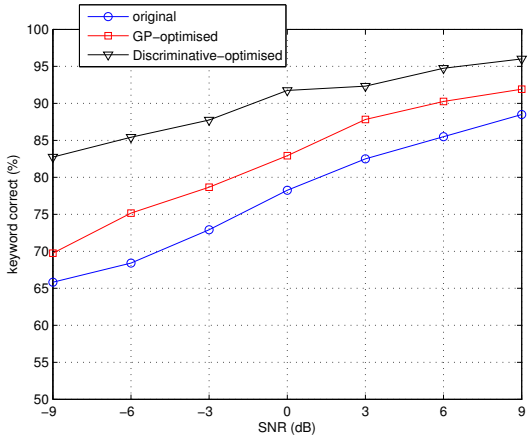

Figure 4: HMM-based intelligibility model estimates for the unprocessed speech, GP-optimised speech and discriminative-optimised speech in stationary noise at a range of SNRs average across the full 600 utterance test set.

speaker) that have not been used during training. We consider three conditions: i/ the unmodified speech, ii/ the GP-optimised speech and iii/ the discriminative-optimised speech. For all experiments we have employed stationary pink noise as the masker. The noise is added to the original speech and the processed speech at the following range of signal-to-noise (SNR) ratios: -9, -6, -3, 0, 3, 6 and 9 dB.

For each noisy utterance the optimal speech spectral shaping parameters, $\mathbf{c}$ are determined using the Nelder-Mead Direct (i.e. 'simplex') Search method. This simple optimisation approach only requires evaluation of the function and does not require derivatives. It was found to converge quickly when the shaping vector $\mathbf{c}$ was initialised to $\mathbf{0}$ so there was no need to use more sophisticated algorithms.

Figure 1 shows an example representation of the glimpse mask of one utterance for the unprocessed speech *(upper panel)*, The GP-based method *(middle panel)*, and finally the discriminative-based method *(lower panel)*. It can be seen that the GP-based method successfully increases the GP from 26% to over 32%. In contrast, the discriminative-based method has nearly the same GP as the unprocessed speech but with a clearly different masking pattern, i.e. to maximise discrimination it was optimal to simply minimise the masking in the GP sense – some frequencies appear to be more informative than others.

These results are consistent across the entire 600 utterance test set. For each utterance the masking patterns generated by the GP and discriminative techniques were compared. Figure 2 shows how the GP score for the GP-optimised technique is always larger and the difference increases as SNR increases. The figure also compares the masks in terms of number of pixels that have changed and it can be seen that the two techniques lead to masks that differ by between 6 and 16% of their extent.

The difference in masking patterns can be understood by looking at the optimal spectral shaping weights for each technique as shown in Figure 3. The figure shows results for male and female speakers separately. It can be seen that the GP technique leads to a spectral shaping that redistributes energy by taking it from the region around channel 10 and adding it around channel 25 to 30. The discriminative approach by contrast at first sight appear to be making much smaller changes, but note these are averages across utterance: the discriminative plot has much larger standard error bars - the smaller average is arising because the discriminative approach leads to a much wider range of strategies. It is also interesting to note that the average effect is subtly different from the GP-optimisation. The very highest frequencies are not boosted.

Finally, we can compare the intelligibilities predicted by the microscopic model by looking at the percentage keywords that are predicted to be recognised correctly (Figure 4). For the original speech the intelligibilities range from 66% to 89% for the original speech. The GP approach boosts these scores to 70% to 92%. For the discriminative approach the scores are increased to 83% to 96%. Note, this HMM-based microscopic model has been previously evaluated in stationary noise masking conditions and has been shown to be a good predictor of listener performance.

5. Conclusions

The paper has proposed a novel technique for optimising the spectral shaping applied by a pre-enhancement system. The approach seeks to maximise intelligibility with respect to a microscopic intelligibility model and is designed to be used in situations where both the speech and noise signal are known apriori and where a statistical model of the speech is available. Compared to approaches that only consider the degree of spectro-temporal masking (e.g. GP optimisation) the new approach finds more varied utterance-dependent tunings of the shaping parameters. Evidence from a microscopic intelligibility model (similar to one that has been validated in the same noise conditions [21]) suggest that the new approach produces a more intelligible result. Listening example are provided on the authors' web site for readers to judge[1].

The current experiments have employed stationary noise and speech material that has come from a corpus that allows construction of very precise speaker-dependent models. These settings may be appropriate as a first approximation in many application settings, but we now wish to generalise the approach to handle more complex data. Current work is examining dynamically varying shaping filters that will allow the enhancement to adapt to changing characteristics of the background. Formal listening tests are also planned to verify the predictions made by the objective intelligibility models.

[1] `http://staffwww.dcs.shef.ac.uk/people/m.aldabel/`

6. References

[1] M. Picheny, N. Durlach, and L. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *Journal of Speech, Language and Hearing Research*, vol. 28, no. 1, p. 96, 1985.

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

[3] ——, "Near end listening enhancement considering thermal limit of mobile phone loudspeakers," in *Proc. Conf. on Elektronische Sprachsignalverarbeitung (ESSV), Aachen, Germany*, 2011, pp. 333–340.

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

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

[6] ——, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech, Portland, USA*, 2012.

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

[8] P. N. Petkov, G. E. Henter, and W. B. Kleijn, "Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 1035–1045, 2013.

[9] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

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

[11] J. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, p. 510, 1993.

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

[13] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, p. 3261, 2008.

[14] H. Brouckxon, W. Verhelst, and B. De Schuymer, "Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments," *signal*, vol. 1, p. 5, 2008.

[15] 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.

[16] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

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

[18] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the Nelder–Mead simplex method in low dimensions," *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[19] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *speech communication*, vol. 45, no. 1, pp. 5–25, 2005.

[20] 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, p. 2421, 2006.

[21] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, vol. 49, no. 5, pp. 402–417, May 2007.