Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the Glimpse Proportion

Cassia Valentini-Botinhao[a,*], Junichi Yamagishi[a], Simon King[a], Ranniery Maia[b]

[a] *The Centre for Speech Technology Research, University of Edinburgh, UK*
[b] *Cambridge Research Laboratory, Toshiba Research Europe Limited, UK*

Abstract

This paper describes speech intelligibility enhancement for hidden Markov model (HMM) generated synthetic speech in noise. We present a method for modifying the Mel cepstral coefficients generated by statistical parametric models that have been trained on plain speech. We update these coefficients such that the Glimpse Proportion – an objective measure of the intelligibility of speech in noise – increases, while keeping the speech energy fixed. An acoustic analysis reveals that the modified speech is boosted in the region 1-4kHz, particularly for vowels, nasals and approximants. Results from listening tests employing speech-shaped noise show that the modified speech is as intelligible as a synthetic voice trained on plain speech whose duration, Mel cepstral coefficients and excitation signal parameters have been adapted to Lombard speech from the same speaker. Our proposed method does not require these additional recordings of Lombard speech. In the presence of a competing talker, both modification and adaptation of spectral coefficients give more modest gains.

Keywords: intelligibility of speech in noise, HMM-based speech synthesis, Mel cepstral coefficients, Glimpse Proportion measure

*Corresponding author. Tel.: +441316511768 / Fax:+441316506626
Email address: C.Valentini-Botinhao@sms.ed.ac.uk (Cassia Valentini-Botinhao)

1. Introduction

In a conversation, humans vary the way they perceive and produce speech according to context. Humans are able to modulate various properties of their speech in order to maintain successful communication in varying contexts, including changes in the listening environment (Picheny et al., 1985; Summers et al., 1988; Lindblom, 1990; Howell et al., 2006; Patel and Schell, 2008; Cooke and Lu, 2010).

For machines, that communicate using speech, to achieve human-like levels of intelligibility in varying contexts, they must also adjust appropriate properties of their speech output. Currently, systems such as Text-To-Speech (TTS) synthesisers are 'deaf' to the context: they speak in the same way, regardless of the listening environment.

In this work we focus solely on automatic strategies to improve speech intelligibility in the presence of additive background noise. In particular we are interested in increasing the intelligibility in noise of synthetic speech generated by a TTS system which employs statistical parametric models – 'Hidden Markov Model (HMM)-based' speech synthesis (Zen et al., 2009). We work under the assumption that the noise signal is available and under the constraint that the signal to noise ratio (SNR) should remain fixed i.e., our intelligibility enhancement method should not merely increase the overall energy level of an utterance. An additional important design constraint is to create intelligible voices without relying on the availability of natural speech produced under matched conditions, since this approach is unlikely to scale to different SNRs and noise types. Our proposed method requires only conventional speech recordings made in quiet conditions, of the type normally used to build TTS systems.

In a quiet listening environment, the intelligibility of state-of-the-art HMM-generated synthetic speech can be as good as that of natural speech (Yamagishi et al., 2008). However, in noisy environments, unmodified synthetic speech tends to reduce in intelligibility to a much greater extent than unmodified natural speech (King and Karaiskos, 2010). By modifying the synthetic speech, either via the statistical models or the acoustic features, it is possible to control the characteristics of the generated speech without the need for new data and so generate synthetic speech that is more intelligible in noise than the natural speech used for training (King and Karaiskos, 2010; Suni et al., 2010; Bonardo and Zovato, 2007). One way to do this is to reproduce some of the acoustic changes observed, in many previous studies, of

natural speech produced in noise: so-called Lombard speech. Another strategy is to use this data through voice conversion techniques (Langner and Black, 2005) and adaptation techniques (Raitio et al., 2011). Research has also been conducted into the generation of hyperarticulated synthetic voices (B. Picart, 2011; Nicolao et al., 2012) that come under the category of clear speech rather than speech specifically produced to counteract the effects of noise.

Lombard speech is speech produced by a talker who is simultaneously listening to noise. It is more intelligible than speech produced in quiet, when each are played to listeners mixed with the same noise and at the same SNR that the talker was experiencing (Summers et al., 1988; Junqua, 1993; Lu and Cooke, 2008). It has also been found that Lombard speech has distinct acoustic differences to speech produced in quiet: overall intensity increases, fundamental frequency increases, flatter spectral tilt (more energy at high frequencies), speaking rate changes (longer vowels / shorter consonants) and formants tend to shift (F1 increased F2 decreased) (Summers et al., 1988; Junqua, 1993; Hansen, 1996; Garnier et al., 2006; Lu and Cooke, 2008). It remains relatively unclear which of these acoustic changes improve intelligibility (and why), or how and to what extent these changes depend on the noise signal. As a consequence, it is non-trivial to use the known properties of Lombard speech to automatically improve the intelligibility of (synthetic) speech in noise.

In this work we present a method to increase the intelligibility of synthetic speech in noise. The method modifies speech automatically according to the known noise characteristics. Rather than using knowledge about speech production in noise (e.g., Lombard speech, as above), we use well-established models of speech perception in noise. Using these models we can obtain reliable estimates of the impact that noise has on the intelligibility of speech; these models also give reliable estimates for modified speech.

Approaches that have been proposed to modify natural speech according to the noise signal include: modification of the local signal-to-noise ratio (SNR) (Sauert and Vary, 2006; Tang and Cooke, 2010) and objective measure based spectral power optimisation using the Speech Intelligibility Index (Sauert and Vary, 2011), a genetic algorithm optimisation for spectral weighting based on the Glimpse Proportion (GP) (Cooke, 2006; Tang and Cooke, 2012) and an optimisation algorithm using a spectro-temporal measure based on a multi stage perceptual model Taal et al. (2012).

In previous work, we have demonstrated that simple changes in the spec-

tral domain (McLoughlin and Chance, 1997) can result in significant gains in intelligibility for HMM-generated synthetic speech across a variety of noise and SNR conditions (Valentini-Botinhao et al., 2011a). On the other hand, changes to fundamental frequency and spectral peaks were not as effective. In the same study we also evaluated a number of objective intelligibility measures, to discover how well they could predict these gains. Taking into consideration both performance and computational complexity, we selected the Glimpse Proportion (GP) measure (Cooke, 2006) as being most suitable for the current task. We then proposed a method to extract cepstral coefficients which maximized the GP measure (Valentini-Botinhao et al., 2012a). Although we obtained intelligibility gains using these extracted cepstral coefficients to train a TTS voice, we also observed distortions in the synthetic speech. Because this method is applied at feature extraction time it requires a new model to be trained for each different noise type (and potentially each SNR) and consequentially cannot handle fluctuating noise. Our solution to this was to modify the generated speech instead (Valentini-Botinhao et al., 2012b), by modifying the Mel cepstral coefficients. In this paper we present the full development of this idea plus a detailed analysis of how the modification actually changes the synthetic speech. We also provide the complete mathematical derivation of the method.

In Section 2 we define the utilized Mel cepstral coefficients and Section 3 we show how the GP measure operates. Section 4 shows in detail how we reformulate GP so that it is completely defined by the Mel cepstral coefficients, then Section 5 describes how to use this approximated measure as part of a method to modify these coefficients. We then present in Sections 6 and 7 convergence and acoustic analyses as well as subjective intelligibility scores for two different listening evaluations. Conclusions are given in Section 8.

2. Mel cepstral coefficients

We can represent the spectrum of speech $H(e^{j\omega})$ by a M-th order Mel cepstral coefficient set $\{c_m\}_{m=0}^{M}$ (Fukada et al., 1992):

$$H(e^{j\omega}) = \exp \sum_{m=0}^{M} c_m e^{-jm\tilde{\omega}} \tag{1}$$

$$\tilde{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{2}$$

where α is the warping factor that controls the frequency scaling.

We can choose α such that $\tilde{\omega}$ spans the frequency axis on a particular scale, such as for instance the Mel scale, creating so-called Mel cepstral coefficients (Fukada et al., 1992). When using the Mel scale warping we can represent the spectrum envelope with fewer coefficients than when using a linear frequency scale, without a loss in quality.

According to eq.(1), the magnitude spectrum is defined by the Mel cepstral coefficients as follows:

$$|H(e^{j\omega})| = \exp \sum_{m=0}^{M} c_m \cos(m\,\tilde{\omega}) \tag{3}$$

3. The glimpse proportion measure

The Glimpse Proportion measure was originally proposed in the context of the Glimpse model for speech perception in noise (Cooke, 2003). The model is based on the ability of humans to obtain information from those time-frequency regions where speech is less masked by noise and therefore less distorted (Cooke, 2003).

The GP measure (Cooke, 2006) is based on this concept: in a noisy environment, humans focus their auditory attention on 'glimpses' of speech that are not masked by noise. Rather than being a correlation, a distance or a ratio, the GP is based on detection: to measure the number of available glimpses of a given speech signal in a given noise, we need the speech and noise signals to be available separately.

The GP correlates well with subjective scores for intelligibility of natural speech in noise (Cooke, 2006). In our own experiments, we also observed similar behaviour for the intelligibility of HMM-generated speech in noise (Valentini-Botinhao et al., 2011b) even when that speech has been modified (Valentini-Botinhao et al., 2011a). In that experiment, we modified parameters such as the fundamental frequency (F_0) and spectral tilt to emulate Lombard speech properties; even under such modifications, GP was a reasonable intelligibility predictor (Valentini-Botinhao et al., 2011a). In all these different scenarios, GP outperformed most other measures in terms of accurate predictions of intelligibility of speech in noise. An attractive property of GP is that its implementation does not require any time delays.

The GP measure is simply the proportion of spectro-temporal regions, so called 'glimpses', where speech is more energetic than noise. To detect such

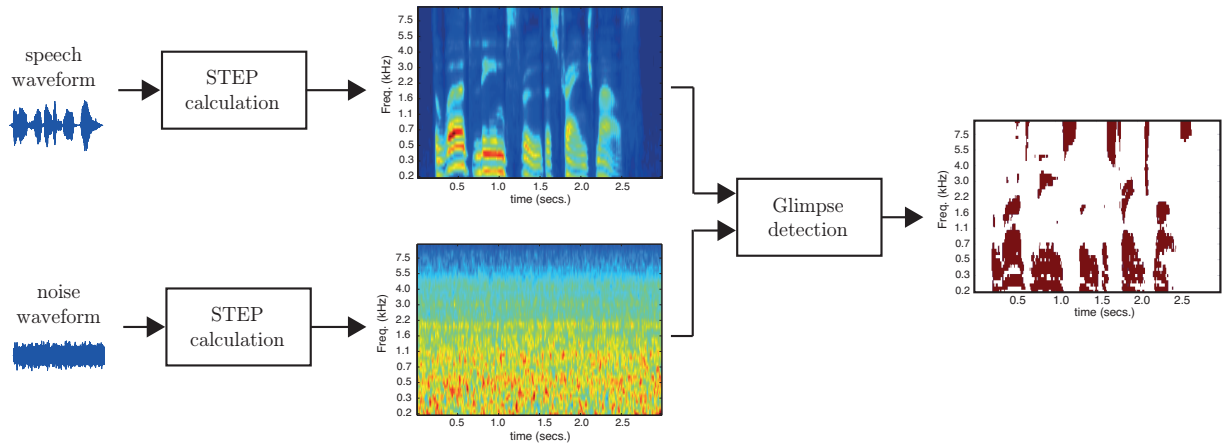


Figure 1: The glimpse proportion measure.

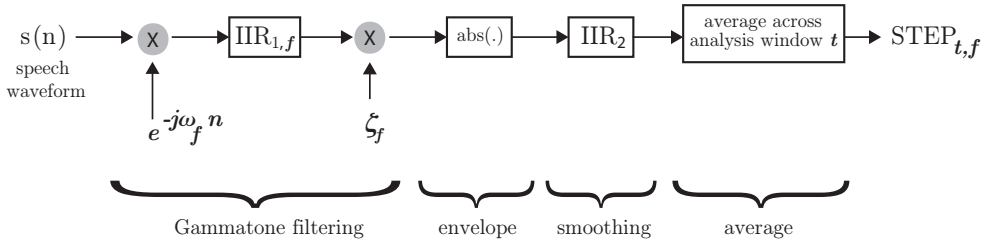


Figure 2: The spectro temporal excitation pattern (STEP) calculation for time frame t and frequency channel f, where IIR_1, f refers to the infinite impulse response (IIR) Gammatone filter, IIR_2 is the smoothing filter and ζ_f is the gain that normalizes the Gammatone filter responses across frequency channels.

glimpses we compare speech and noise using the spectro-temporal excitation pattern (STEP) representation as shown in Figure 1. To represent a signal in terms of STEP – see Figure 2 – we first decompose its waveform into different frequency channels using a Gammatone filterbank whose central frequencies are linearly spaced on the Equivalent Rectangular Bandwidth (ERB) scale (Moore and Glasberg, 1996). For each channel, the temporal envelope is extracted with an absolute value operation, smoothed with a low pass filter and then averaged across limited time intervals. A glimpse is detected in a time frequency region when the speech STEP value in that region is higher than the noise value.

The parameters that define the GP measure are: the range of the Gammatone filters' centre frequencies (100-7500 Hz), the number of Gammatone

filters N_f (55 filters), the temporal integration time for the smoothing filter (8 ms), the size of the time frame (30 ms) and its period (10 ms).

4. Proposed GP approximation

In this section we show how we can approximate the GP measure so that it is completely defined by the short term magnitude spectrum of speech and consequently by the sequence of Mel cepstral coefficients.

To obtain a closed and differentiable formula that relates spectral parameters to the GP measure we make the following approximations and correspondences:

- the input signals are no longer the signal waveforms of speech and noise but the short term magnitude spectrum calculated from the short-time Mel cepstral coefficients of speech and from the short-time discrete Fourier transform of noise (approximation)
- the previous approximation implies that all operations are carried out in the frequency domain rather than the time domain (correspondence)
- the filtering operations in the time domain are replaced by multiplications in the frequency domain with a truncated version of the frequency responses of the infinite impulse response filters (approximation due to the truncation)
- the absolute value in the time domain is replaced by a power operation that can be represented in the frequency domain as the circular convolution operation (approximation)
- the hard threshold detection of glimpses is replaced by a soft decision threshold defined by a sigmoid function (generalization)

The proposed approximated GP measure is then given by:

$$GP = \frac{100}{N_f N_t} \sum_{t=1}^{N_t} \sum_{f=1}^{N_f} \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \tag{4}$$

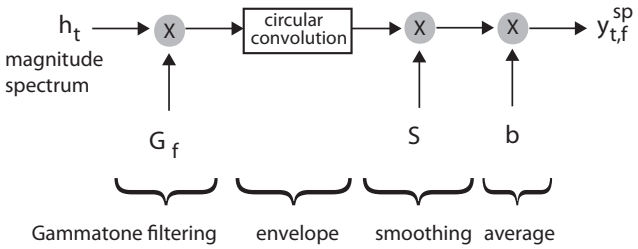


Figure 3: Proposed approximation for STEP calculation (Valentini-Botinhao et al., 2012a).

where the following scalars are defined as:

$y^{sp}_{t,f}$	STEP approximation for speech at analysis window t and frequency channel f
$y^{ns}_{t,f}$	STEP approximation for noise at analysis window t and frequency channel f
N_t	number of time frames
N_f	number of frequency channels
t	analysis window index
f	frequency channel index
$\mathcal{L}(.)$	a logistic sigmoid function defined as:

$$\mathcal{L}(x) = \frac{1}{1 + e^{-\eta x}} \tag{5}$$

where η defines the slope of the curve.

The STEP approximation as seen in Figure 3 is given by:

$$y^{sp}_{t,f} = \frac{1}{N}(\boldsymbol{G}_f \boldsymbol{h}_t \circledN \boldsymbol{G}_f \boldsymbol{h}_t)^\top \boldsymbol{S}\, \boldsymbol{b} \tag{6}$$

where:

N number of frequency bins of the spectrum

$\boldsymbol{h}_t = \left[\, |H_t(\omega_1)| \;\ldots\; |H_t(\omega_N)| \,\right]^\top$
vector Nx1 - magnitude spectrum of windowed speech signal $\boldsymbol{s}$ at analysis window t

$\boldsymbol{G}_f = \text{diag}\left(\left[\, g_{f,1} \;\ldots\; g_{f,N} \,\right]\right)$
matrix NxN - diagonal matrix, diagonal contains the Gammatone filter frequency response for frequency channel f

$\boldsymbol{S} = \text{diag}\left(\left[\, \mathsf{s}_1 \;\ldots\; \mathsf{s}_N \,\right]\right)$
matrix NxN - diagonal matrix, diagonal contain the frequency response of the smoothing filter

$\boldsymbol{b} = \left[\, b_1 \;\ldots\; b_N \,\right]$
vector Nx1 - coefficients of average filter

$\textcircled{\scriptsize N}$ circular convolution operation dimension N

5. GP-based Mel cepstral modifications

In this section we show how to modify a sequence of Mel cepstral coefficients generated by a statistical model, such that the GP measure increases.[1]

5.1. Cost function

To increase glimpses in a certain analysis window t we first define the following cost function:

$$GP_t(\boldsymbol{c}_t) = \frac{100}{N_f} \sum_{f=1}^{N_f} \mathcal{L}(y_{t,f}^{sp}(\boldsymbol{c}_t) - y_{t,f}^{ns}) \tag{7}$$

[1] Although the formulation of the problem allows for the extension to other types of spectral parametrization such as the Mel Generalized Cepstral coefficients (MGC) (Koishida et al., 1996) we can not guarantee that the synthesis filter created from such modified MGCs is stable. Stability is always guaranteed for any value of Mel cepstral coefficients. To modify the MGC parameters it would be necessary to first transform them into a representation where stability is easily ensured like the MGC-LSP as proposed in (Koishida et al., 2000).

The optimal spectral parameter vector $\boldsymbol{c}_t = [c_{t,1} \quad c_{t,2} \ldots \quad c_{t,M}]^\top$ would then be given by:

$$\boldsymbol{c}_t^* = \operatorname{argmax} GP_t(\boldsymbol{c}_t) \tag{8}$$

5.2. Steepest descent solution

The update equation for Mel cepstral coefficients using steepest descent is given by:

$$\boldsymbol{c}^{(i+1)} = \boldsymbol{c}^{(i)} + \mu \boldsymbol{\Delta} c^{(i)} \tag{9}$$

$$= \boldsymbol{c}^{(i)} + \mu \boldsymbol{\nabla} GP_t^{(i)}(\boldsymbol{c}_t) \tag{10}$$

where $\boldsymbol{\Delta} c^{(i)}$ is the Mel cepstral coefficient increment in iteration i, $\boldsymbol{\nabla} GP_t^{(i)}(\boldsymbol{c}_t)$ is the gradient of the function defined in eq.(7) in iteration i and μ is the stepsize.

From now on we will drop the iteration index (i) and the argument $(\boldsymbol{c}_t)$ for clarity. We can find the gradient vector as follows:

$$\boldsymbol{\nabla} GP_t = \frac{\partial GP_t}{\partial \boldsymbol{c}_t} = \frac{100}{N_f} \sum_{f=1}^{N_f} \frac{\partial \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})}{\partial \boldsymbol{c}_t} \tag{11}$$

$$= \frac{100}{N_f} \sum_{f=1}^{N_f} \frac{\partial \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})}{\partial y_{t,f}^{sp}} \frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{c}_t} \tag{12}$$

The first term in the summation of eq.(12) can be written as:

$$\frac{\partial \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})}{\partial y_{t,f}^{sp}} = \eta\, \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \big[1 - \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})\big] \tag{13}$$

The second term in the summation of eq.(12) is given by:

$$\frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{c}_t} = \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{c}_t} \frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{h}_t} \tag{14}$$

The first term on the right of eq.(14) is a matrix of dimension MxN defined as:

$$\boldsymbol{H}_{c_t} \equiv \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{c}_t} = \begin{bmatrix} \frac{\partial |H_t(\omega_1)|}{\partial c_{t,1}} & \frac{\partial |H_t(\omega_2)|}{\partial c_{t,1}} & \cdots & \frac{\partial |H_t(\omega_N)|}{\partial c_{t,1}} \\ \frac{\partial |H_t(\omega_1)|}{\partial c_{t,2}} & \frac{\partial |H_t(\omega_2)|}{\partial c_{t,2}} & \cdots & \frac{\partial |H_t(\omega_N)|}{\partial c_{t,2}} \\ \vdots & & & \\ \frac{\partial |H_t(\omega_1)|}{\partial c_{t,M}} & \frac{\partial |H_t(\omega_2)|}{\partial c_{t,M}} & \cdots & \frac{\partial |H_t(\omega_N)|}{\partial c_{t,M}} \end{bmatrix}$$

When the spectrum is modelled by Mel cepstral coefficients as in eq.(1), the elements of this matrix are defined as:

$$\{\boldsymbol{H}_{c_t}\}_{m,k} = \frac{\partial |H_t(\omega_k)|}{\partial c_{t,m}} = |H_t(\omega_k)| \cos(m\,\tilde{\omega}_k) \tag{15}$$

where k is the index for the spectrum frequency bin and m as defined previously is the index for the Mel cepstral coefficient sequence. The second term of eq.(14) depends on the definition of the STEP approximation in eq.(6) and it is then given by:

$$\begin{aligned}
\frac{\partial y^{sp}_{t,f}}{\partial \boldsymbol{h}_t} &= \frac{\partial \boldsymbol{l}_{t,f}}{\partial \boldsymbol{h}_t}\,\frac{\partial y^{sp}_{t,f}}{\partial \boldsymbol{l}_{t,f}} && (16)\\
&= \frac{1}{N}\frac{\partial \boldsymbol{l}_{t,f}}{\partial \boldsymbol{h}_t}\,\boldsymbol{S}\,\boldsymbol{b} && (17)\\
&= \frac{1}{N}\frac{\partial \boldsymbol{G}_f\boldsymbol{h}_t}{\partial \boldsymbol{h}_t}\,\frac{\partial \boldsymbol{l}_{t,f}}{\partial \boldsymbol{G}_f\boldsymbol{h}_t}\,\boldsymbol{S}\,\boldsymbol{b} && (18)\\
&= \frac{1}{N}\boldsymbol{G}_f\,\frac{\partial \boldsymbol{l}_{t,f}}{\partial \boldsymbol{G}_f\boldsymbol{h}_t}\,\boldsymbol{S}\,\boldsymbol{b} && (19)\\
&= \frac{1}{N}\boldsymbol{G}_f\,(2\,\boldsymbol{\Gamma}_N \,\textcircled{N}\, \boldsymbol{G}_f\boldsymbol{h}_t)\,\boldsymbol{S}\,\boldsymbol{b} && (20)
\end{aligned}$$

where $\boldsymbol{l}_{t,f} = (\boldsymbol{G}_f\boldsymbol{h}_t \,\textcircled{N}\, \boldsymbol{G}_f\boldsymbol{h}_t)$ and $\boldsymbol{\Gamma}_N$ is the identity matrix of dimension N. The operation $(\boldsymbol{\Gamma}_N \,\textcircled{N}\, \boldsymbol{G}_f\boldsymbol{h}_t)$ defines a matrix NxN of the following form:

$$\begin{bmatrix}
\boldsymbol{e}_1 \,\textcircled{N}\, (\boldsymbol{G}_f\boldsymbol{h}_t)^\top \\
\boldsymbol{e}_2 \,\textcircled{N}\, (\boldsymbol{G}_f\boldsymbol{h}_t)^\top \\
\vdots \\
\boldsymbol{e}_N \,\textcircled{N}\, (\boldsymbol{G}_f\boldsymbol{h}_t)^\top
\end{bmatrix}$$

where $\boldsymbol{e}_n$ is the n-th column of the identity matrix $\boldsymbol{\Gamma}_N$.

Connecting eqs.(13), (15) and (20), the gradient vector is given by:

$$\begin{aligned}
\boldsymbol{\nabla} GP_t = \frac{100}{N_f N}\sum_{f=1}^{N_f} &\eta\,\mathcal{L}(y^{sp}_{t,f} - y^{ns}_{t,f})\big[1 - \mathcal{L}(y^{sp}_{t,f} - y^{ns}_{t,f})\big] \\
&\boldsymbol{H}_{c_t}\boldsymbol{G}_f\,(2\,\boldsymbol{\Gamma}_N \,\textcircled{N}\, \boldsymbol{G}_f\boldsymbol{h}_t)\,\boldsymbol{S}\,\boldsymbol{b}
\end{aligned} \tag{21}$$

5.3. Energy normalization

In this section we explain how to reformulate the optimization problem in order to keep the overall energy of speech constant. For clarity, we drop the index t in the equations and use the continuous representation of the spectrum $H(e^{j\omega})$.

Let us first define the quantity we refer to here as the overall energy of speech in a certain time frame:

$$\sum_{n=0}^{N-1} |s(n)|^2 = \psi \tag{22}$$

From Parseval's theorem we have that:

$$\sum_{n=0}^{N-1} |s(n)|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S(e^{j\omega})|^2 \, d\omega = \psi \tag{23}$$

where $S(e^{j\omega})$ is the discrete time Fourier transform of time signal $s(n)$.

This can be related to the spectral envelope $H(e^{j\omega})$ and the frequency representation $E(e^{j\omega})$ of the excitation signal:

$$\psi = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})E(e^{j\omega})|^2 \, d\omega \tag{24}$$

We can assume that $|E(e^{j\omega})|$ is constant over the frequency domain for both voiced and unvoiced segments. For voiced speech segments this is true if the size of the analysis window is set to two pitch periods and for unvoiced segments this is true because at these segments the excitation signal is white noise. Under this assumption and considering that the cepstral extraction method does not modify the excitation signal we can assume that in order to keep the energy in the time domain constant it is sufficient to keep the following constant:

$$\psi = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \, d\omega \tag{25}$$

The maximization of the glimpse function as given in eq.(5.1) should then be solved subject to the above constraint. Solving a nonconvex optimization problem is however a hard task. One feasible solution is to perform at each iteration of the Steepest Descent method an energy normalization operation

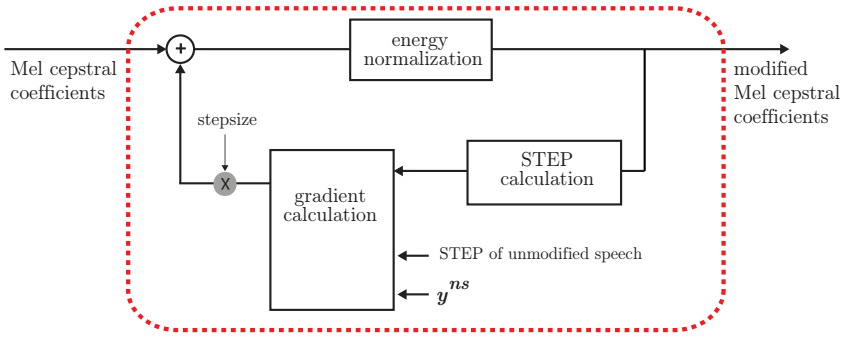


Figure 4: GP-based Mel cepstral coefficient modification using steepest descent with energy normalization.

and alter the objective function and consequentially the gradient vector accordingly. Figure 4 shows this solution. To explain how the gradient should be modified we first need to define the operation that normalizes the energy of the spectrum.

The following operation modifies the spectrum $|H(e^{j\omega})|$ with overall energy ψ so that the resulting spectrum $|H'(e^{j\omega})|$ has an overall energy equal to ψ':

$$|H'(e^{j\omega})| = \frac{|H(e^{j\omega})|}{\sqrt{\frac{1}{\psi'}\int_{-\pi}^{\pi}|H(e^{j\omega})|^2 d\omega}} = \frac{|H(e^{j\omega})|}{\sqrt{\frac{\psi}{\psi'}}} \tag{26}$$

In order to modify the gradient we need to calculate the impact of this operation in the Mel cepstral coefficient domain. The normalization operation transforms a set of Mel cepstral coefficients c_m that model the spectrum $|H(e^{j\omega})|$ with overall energy ψ, into parameters c'_m that model a spectrum

$|H'(e^{j\omega})|$ with overall energy equal to ψ' in the following way:

$$|H'(e^{j\omega})| = \frac{|H(e^{j\omega})|}{\sqrt{\frac{\psi}{\psi'}}} = \frac{\exp \sum_{m=0}^{M} c_m \cos(m\,\tilde{\omega})}{\exp\left[\log \sqrt{\frac{\psi}{\psi'}}\right]} \tag{27}$$

$$= \exp\left[\Big(\sum_{m=0}^{M} c_m \cos(m\,\tilde{\omega})\Big) - 0.5 \log\left(\frac{\psi}{\psi'}\right)\right] \tag{28}$$

$$= \exp \sum_{m=0}^{M} c'_m \cos(m\,\tilde{\omega}) \tag{29}$$

The energy-normalized Mel cepstral coefficients c'_m are then given by:

$$c'_m = \begin{cases} c_0 - 0.5 \log\left(\frac{\psi}{\psi'}\right) & m = 0 \\ c_m & m \neq 0 \end{cases} \tag{30}$$

Only the c_0 coefficient changes, so we can write the energy normalized magnitude spectrum as:

$$|H'(e^{j\omega})| = |K'||D(e^{j\omega})| \tag{31}$$

where $K' = \exp(c'_0)$ and $D(e^{j\omega}) = \exp \sum_{m=1}^{M} c_m e^{-jm\,\tilde{\omega}}$.

If ψ is equal to ψ', i.e. the energy-normalization operation has no impact on the spectrum, we can see that c'_m is equal to c_m. The only term in the gradient vector $\boldsymbol{\nabla} GP$ that needs to be adjusted is the one given by eq.(15). To show how this term changes we adopt the discrete representation $H(\omega_1), \ldots, H(\omega_N)$ of the spectrum. Eq.(25) is then approximated to:

$$\psi = \sum_{j=1}^{N} |H(\omega_k)|^2 \tag{32}$$

With the energy normalization operation, the derivative in eq.(15) becomes:

$$\frac{\partial|H'(\omega_k)|}{\partial c_m} = \frac{\partial|K'||D(\omega_k)|}{\partial c_m} \tag{33}$$

$$= \frac{\partial|K'|}{\partial c_m}|D(\omega_k)| + |K'|\frac{\partial|D(\omega_k)|}{\partial c_m} \tag{34}$$

$$= |K'|\frac{\partial c_0'}{\partial c_m}|D(\omega_k)| + |K'||D(\omega_k)|\cos(m\,\tilde{\omega}_k) \tag{35}$$

$$= |H'(\omega_k)|\frac{\partial c_0'}{\partial c_m} + |H'(\omega_k)|\cos(m\,\tilde{\omega}_k) \tag{36}$$

$$= |H'(\omega_k)|\Big(\frac{\partial c_0'}{\partial c_m} + \cos(m\,\tilde{\omega}_k)\Big) \tag{37}$$

The derivative term in the previous equation is given by:

$$\frac{\partial c_0'}{\partial c_m} = \frac{\partial c_0}{\partial c_m} - 0.5\,\frac{\psi'}{\psi}\,\frac{1}{\psi'}\,\frac{\partial \psi}{\partial c_m} \tag{38}$$

$$= \frac{\partial c_0}{\partial c_m} - \frac{1}{\psi}\sum_{l=1}^{N}|H(\omega_l)|^2\cos(m\,\tilde{\omega}_l) \tag{39}$$

$$= \frac{\partial c_0}{\partial c_m} - \frac{1}{\psi'}\sum_{l=1}^{N}|H'(\omega_l)|^2\cos(m\,\tilde{\omega}_l) \tag{40}$$

$$\frac{\partial c_0'}{\partial c_m} = \begin{cases} 0.0 & m = 0 \\ -\frac{1}{\psi'}\sum_{l=1}^{N}|H'(\omega_l)|^2\cos(m\,\tilde{\omega}_l) & m \neq 0 \end{cases} \tag{41}$$

The derivative in eq.(15) becomes then:

$$\frac{\partial|H'(\omega_k)|}{\partial c_m} = \begin{cases} |H'(\omega_k)| & m = 0 \\ |H'(\omega_k)|\Big(\cos(m\,\tilde{\omega}_k) & \\ -\frac{1}{\psi'}\sum_{l=1}^{N}|H'(\omega_l)|^2\cos(m\,\tilde{\omega}_l)\Big) & m \neq 0 \end{cases} \tag{42}$$

Using this new gradient calculation, and normalising the speech energy at each iteration, guarantees that the energy of the speech signal is fixed during gradient descent optimization. Because the optimization is performed per analysis window, the energy of each window will not change, meaning that there is no reallocation of energy across windows and that the maximisation of the GP is bounded by the amount of energy initially available in the analysis window.

5.4. Distortion control

A detection based measure like the GP or a ration based measure like the SNR predicts the effect of additive distortions by comparing the levels of speech and the distortion (in this case noise), not requiring any reference undistorted speech signal. These measures can not predict the effect that modifying speech has on the intelligibility of the noisy mixture. An issue we face then when using the GP measure as an optimization criterion on its own is the need to limit the distortions caused by the modifications. Recent research on improving the GP measure so it can predict intelligibility of modified speech is described in Tang et al. (2013). Our current work is however based on the original measure (Cooke, 2006).

To define an audible distortion, we use the Euclidian distance between the STEP representations of modified and unmodified speech. Including this as an explicit constraint is unfortunately rather cumbersome, so instead we use it as a stopping criterion. We also hypothesize that limiting the frequency resolution of the modifications should generate fewer distortions. This is implemented simply by setting the gradient vector for higher dimensions to zero, and so the method modifies only the first few Mel cepstral coefficients, which represent the coarse properties of the spectrum.

6. First evaluation

In this section we present the details of how the statistical parametric models were built, an analysis of convergence, an acoustic analysis, then the design and results of our first listening experiment. In this first experiment we also test the idea of restricting the frequency resolution of the modifications by updating only the first few Mel cepstral coefficients.

6.1. Voice building

We used two different datasets recorded by the same British male speaker: normal (plain, read-text) speech data and Lombard speech. The Lombard speech was recorded while speech-modulated noise (modulated by the speech from a different male speaker (Dreschler et al., 2001)) was played over headphones at an absolute value of 84 dBA.

Table 1 presents the eight different voices we built for this evaluation. The baseline unmodified voice N was created from a high quality average voice model adapted to 2803 sentences of the normal speech database (three hours of material). Building a speaker-dependent voice was feasible however

Voice	Adaptation	Modification
N	-	-
N-M59	-	all coefficients
N-M10	-	first 10 coefficients
N-M2	-	first 2 coefficients
N-L	only spectrum	-
L	all features	-
L-E	all and extrapolated	-
L-E-M2	all and extrapolated	first 2 coefficients

Table 1: Voices built for the evaluation. All voices were trained using plain natural speech. The "adaptation" column specifies whether each model was adapted to Lombard speech data (noting that "all features" includes duration and excitation parameters). The "modification" column specifies whether the proposed GP-based Mel cepstral modifications were performed on the parameters generated by the models.

the normal speech dataset was not sufficiently phonetically balanced due to the reading material used for the recordings and hence we have decided to use the adaptive approach (Yamagishi et al., 2009). The modified voices N-M59, N-M10 and N-M2 were created from voice N by modifying all, just the first ten (c_1 until c_{10}), or just the first two (c_1 and c_2) Mel cepstral coefficients using the proposed method, as described in the previous section.

We built the other set of voices N-L, L, L-E and L-E-M2 using the Lombard speech portion of the database in addition. Lombard voice L was built by further adapting all parameters (duration, excitation, spectral) of voice N using 780 sentences from the Lombard speech dataset (53 minutes). The reason for not building a voice only with the Lombard dataset was again the lack of phonetic balance in the dataset. Voice N-L was also created from voice N by adapting this time only the Mel cepstral coefficients (i.e., spectral model parameters) to the Lombard data. Voices L-E and L-E-M2 are versions of voice L where we extrapolated the adaptation in all dimensions at an extrapolation factor of 1.2 for Mel cepstral coefficients and 1.35 for duration (voice L-E), and then further modified the two first Mel cepstral using the proposed method (voice L-E-M2).

We trained and adapted the models using the described data sampled at a rate of 48 kHz. We extracted the following acoustic features: 59 Mel cepstral coefficients ($\alpha = 0.77$), Mel scale F0 and 25 aperiodicity energy bands extracted using STRAIGHT (Kawahara et al., 1999). We used a hidden

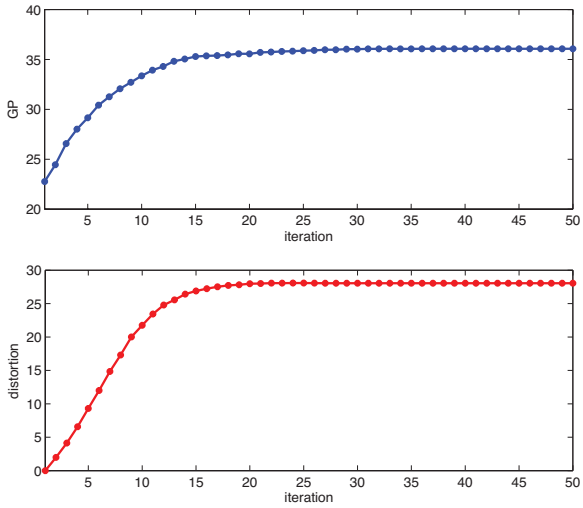


Figure 5: Convergence of (top) GP and (bottom) distortion, averaged over one sentence. Distortion is measured as the percentage increase in the Euclidian distance between the STEP representation of original and modified spectrum.

semi-Markov model as the acoustic model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values, with one stream for the spectrum, three streams for F0 and one for the band-limited aperiodicity. We applied the Global Variance method (Toda and Tokuda, 2007) to compensate for the over smoothing effect caused by the statistical nature of the acoustical modelling.

For the GP-based Mel cepstral modifications we set the following values for the STEP calculation: 55 Gammatone filters with centre frequencies covering the range of 50-7500 Hz (because the noise signal used for testing was sampled at 16 kHz, and so the audio bandwidth was 8 kHz), 8 ms of temporal integration time for the smoothing filter and frame length and period of 30 and 10ms. For the steepest descent optimization we used a normalized step size defined at each iteration i as $\mu^{(i)} = \mu / ||\nabla GP_t^{(i)}||$ (where $\mu = 0.4$ for N-M59 and $\mu = 0.8$ for N-M10 and N-M2). As stopping criteria we use both error convergence and a maximum threshold set to 10 % of relative increase in distortion. We define distortion here as the Euclidian distance between the original and the modified STEP representation of speech.

Voice	Duration (secs.)	Pauses (secs.)	F_0 mean (Hz)	Spectral tilt (dB/oct.)
N	2.11	0.16	104.5	-2.24
N-M2				-1.88
L	2.80	0.19	145.0	-1.70

Table 2: Acoustic properties observed in normal N, modified N-M2 and lombard L voices average across the whole set of sentences used in the listening test (Valentini-Botinhao et al., 2012b).

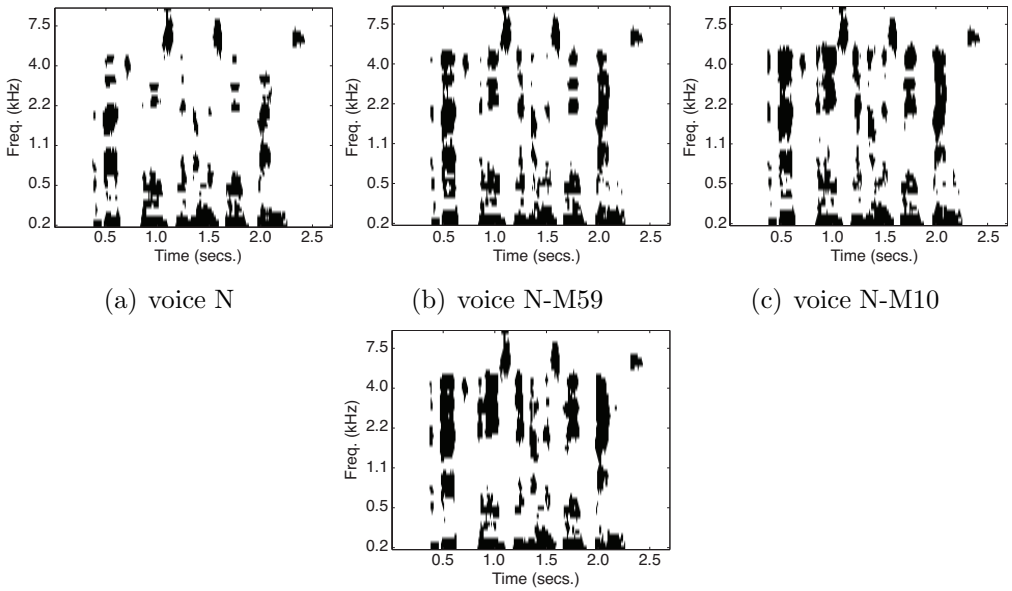
6.2. Convergence analysis

Figure 5 shows the convergence of the GP and distortion values. We can see that, as GP increases, distortion also increases as expected, and that the algorithm is well-behaved (i.e., it converges to a stable value within a reasonable number of iterations). The algorithm is frame-based, meaning that the stopping criteria are applied on a per-frame basis. For individual frames, the convergence is somewhat less smooth-looking than that illustrated in the figure. On average, 5 iterations are sufficient to meet one or other of the stopping criteria for each frame, and more often than not it is the distortion criterion that is met.

6.3. Acoustic analysis

We now examine the impact of the modification at sentence and phone unit levels in terms of GP values and the long term average spectrum (LTAS). The LTAS is calculated as the power spectral density averaged across time frames of 10 ms length and 50 % overlap. This averaged representation is then presented in (dB). First we present a broad analysis across the whole set of sentences used in the listening experiment. Table 2 shows the average duration of speech and pauses, average F_0 and average spectral tilt across all sentences used in the listening test for the normal (N), modified (N-M2) and Lombard (L) voices. As expected, the Lombard voice produces sentences with longer duration and longer pauses, greatly increased F_0 mean and flatter spectral tilt. The modified voice N-M2 produces speech with flatter spectral tilt, though not to the same extent as the Lombard voice.

For a more detailed inspection of the proposed method in operation, Figure 6 shows the glimpses (in black) detected in the presence of speech-shaped noise at -4 dB SNR for (from left to right) a sentence generated by the unmodified voice N and the modified voices N-M59, N-M10 and N-M2. The



(a) voice N (b) voice N-M59 (c) voice N-M10

(d) voice N-M2

Figure 6: Glimpses detected on the STEP time-frequency representation in speech-shaped noise at a SNR of -4 dB for a sentence generated by (a) unmodified voice N and modified voices (b) N-M59 (c) N-M10 and (d) N-M2

glimpses are shown in the STEP domain. We can see that the glimpsed regions become larger and that new glimpses start to appear when we modify all, just the first ten and the first two Mel cepstral coefficients. We also see that when we modify fewer coefficients, the new glimpses tend to be in more coherent regions, creating larger glimpses rather than scattered small glimpses. This is an expected and desired result of modifying only those coefficients that define the coarse shape of the log magnitude spectrum.

Figure 7 shows the GP value for each frame as defined in eq.(7) for the same sentence as shown in Figure 6, generated by the unmodified voice N, the modified voice N-M2 and the Lombard adapted voice L. We observe that, although the number of glimpses on average increases, the increase in glimpses differs between segments. Since the noise that was driving this modification is stationary, this variation comes from the speech signal itself: the different spectral shapes of the various phonetic units will result in fewer or greater numbers of glimpses. In this example sentence, the number of glimpses hardly increases in fricatives and stops, whereas the most substantial

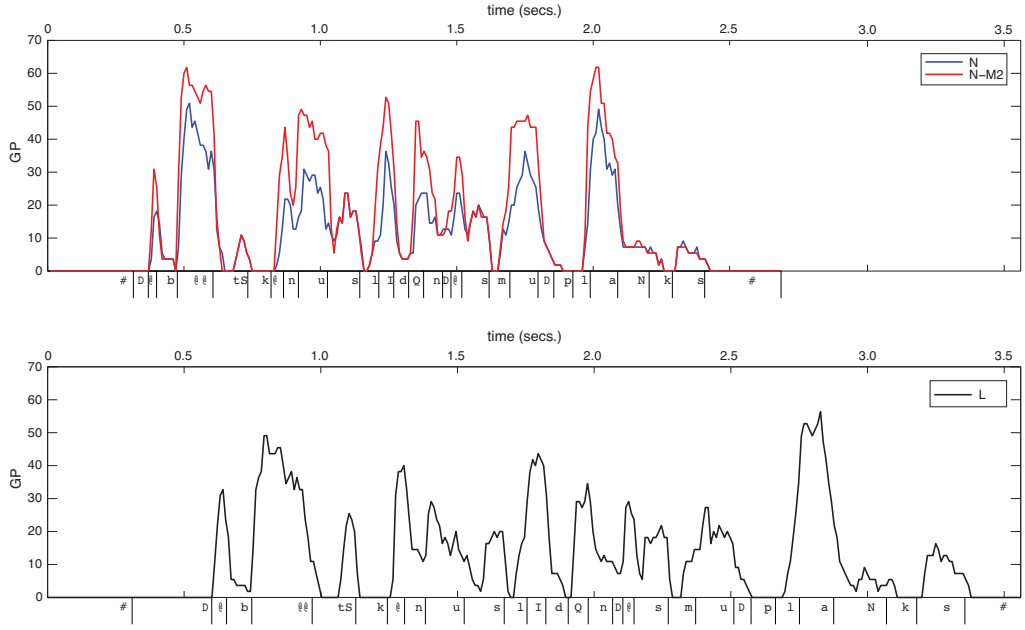


Figure 7: The GP measure across the different frames of the sentence *"The birch canoe slid on the smooth planks"* generated by the original unmodified voice N versus the modified voice N-M2 (top) and the Lombard adapted voice L (bottom), in the presence of speech-shaped noise at −4 dB. The horizontal axis gives the phone segmentation in the combilex phoneset.

increases happen in vowels and nasals. This does not mean that fricatives and stops are not being modified though, but does mean that the proposed method fails to create more glimpses of them for the listener. Although we are not aiming to recreate the Lombard effect, we present the curve obtained from the voice L in the bottom plot of Figure 6. Compared to the GP gains obtained by voice N-L over voice N, the voice L has smaller GP gains during vowels while fricatives' GP values are slightly higher.

For a further detailed analysis we computed the gain in (dB) of the LTAS of voice N-M2 over and above the LTAS of the original unmodified voice N, averaged across all test sentences, for speech-shaped noise at −4 dB. Figure 8 shows the overall pattern of spectral gain at a sentence level and Figures 9 and 10 present the gain calculated for different phonetic classes averaged over all tokens of that class in the test set.

From Figure 8, we observe that, compared to voice N, voice N-M2 exhibits enhanced energy in the region of 1-4 kHz and attenuated below 1 kHz. One

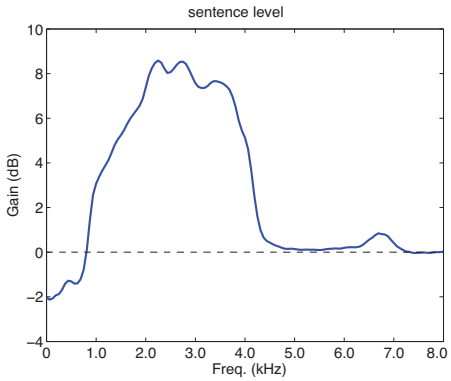


Figure 8: Gain in (dB) of the LTAS of voice N-M2 over the LTAS of unmodified voice N calculated (for speech-shaped noise) at a sentence level and averaged across the set of senteces used in the listening test.

clear observation we can make when comparing the gains for specific phone classes as displayed in Figures 9 and 10 is that the curves as well as the gain values vary substantially across different phonetic classes. In the first group (vowels, nasals and approximants) the gains are at least five times larger than those obtained for the second group (fricatives, affricates and stops). This is a consequence of the shapes and values of the unmodified speech LTAS for these classes.

From the gain curves of the first group displayed in Figure 9 we can see a similar pattern across vowels, nasals and approximants: a large enhancement varying from 8 to 12 dB in the frequency region between approximatively 800 Hz (this number varies across the different classes) and 5 kHz as well as a apparent attenuation of around 2 dB for the lower frequency region. For both vowels and approximants we see also a clear gain region between 5-8 kHz that is separated by a gain valley at approximately 5 kHz. The shapes of these gain curves follow the shape of the LTAS of these phonetic classes, for instance we can see a bump from 5-8kHz in the vowels and approximants. The nasals are the units that are most strongly enhanced reaching a maximum of 12 dB gain which can be explained by the fact that they seem to be highly energetic with an even less flat spectrum than the other sounds.

A similar trend for vowels, nasals and liquids can be seen in a study performed on Lombard speech of 5 male Spanish native speakers Castellanos et al. (1996) although interestingly we did not find this trend in the Lombard database that we recorded from our speaker.

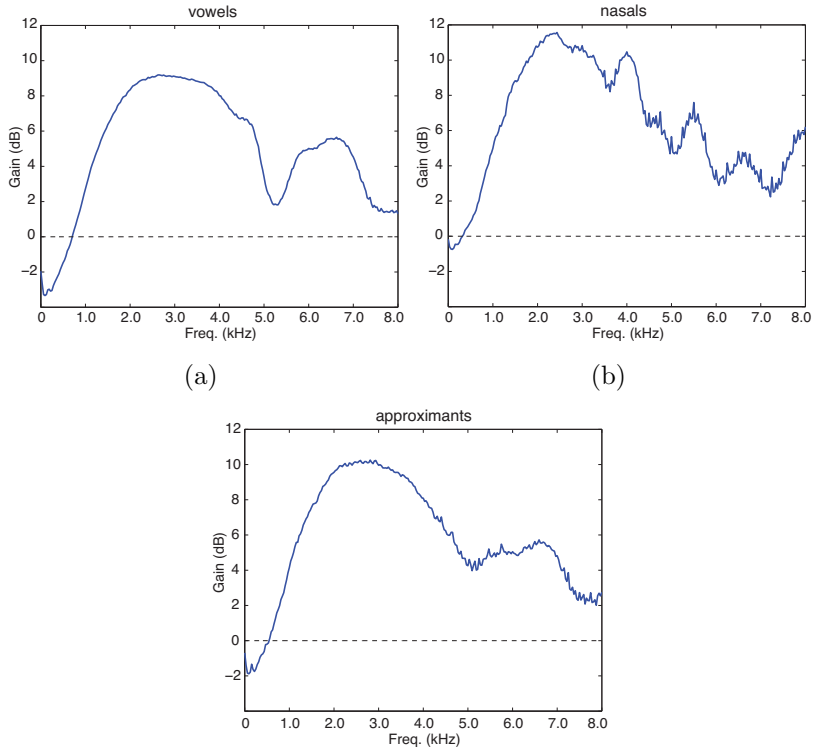


Figure 9: Gain in (dB) of the LTAS of voice N-M2 over the LTAS of unmodified voice N calculated (for speech-shaped noise) averaged across (a) vowels (b) nasals (c) approximants.

The gains obtained for the other class (stops, fricative and affricates) as shown in Figure 10 are, as previously stated, much smaller. For both stops and fricatives an average maximum of 2 dB increase was found and the region most enhanced is between 1-5 kHz as seen for the other group. The affricates show even lower gains and narrow enhanced regions between 1-3 kHz with a valley around 2 kHz.

6.4. Listening experiments design

In this listening test we evaluated the intelligibility of the eight different synthetic voices listed in Table 1 mixed with two noises: speech-shaped (ssn) and speech from a single competing female speaker (cs). To obtain similar intelligibility scores across each noise and to avoid ceiling effects, we mixed each noise at two different SNRs: -4 dB for ssn and -14 dB for the cs. These

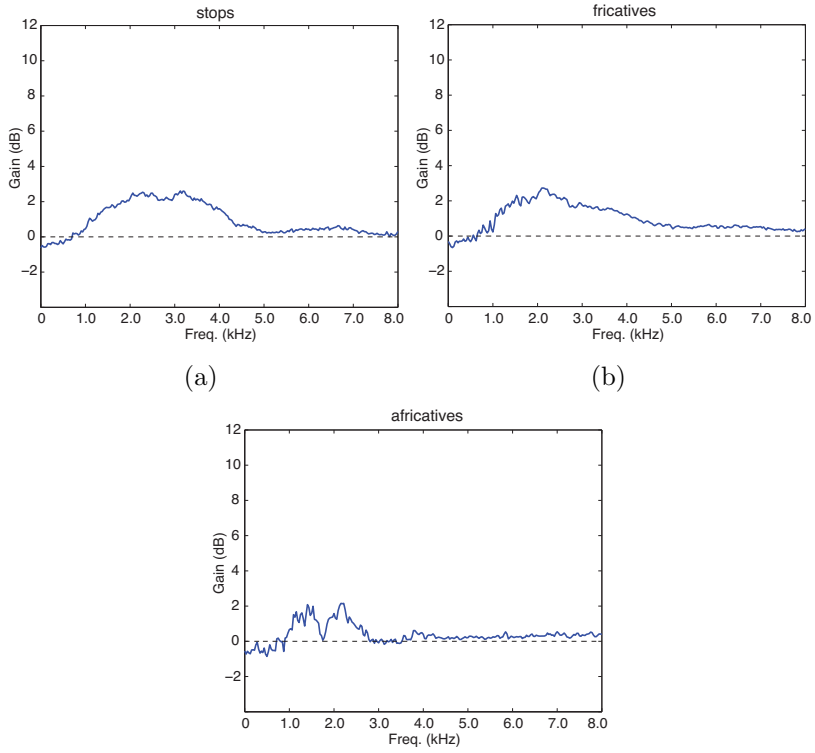


Figure 10: Gain in (dB) of the LTAS of voice N-M2 over the LTAS of unmodified voice N calculated (for speech-shaped noise) averaged across (a) stops (b) fricatives and (c) affricates sounds.

SNRs were chosen to give approximately 50 % word accuracy for natural speech of the same speaker with the same material (Cooke et al., 2013).

32 native English speakers listened to the noisy samples over headphones in soundproof booths. Each participant typed in what he or she heard for a total of six different sentences per condition, i.e., voice and noise type (16 conditions). Each sentence could only be played once and the same sentence was never played again in the same listening test. We used the first ten sets of the Harvard sentences (IEEE, 1969). The Harvard sentences are a group of 720 sentences organized in sets of 10, where each set is designed to be phonetically-balanced. The sentences are also more representative of everyday speech than the semantically unpredictable sentences used in other TTS intelligibility listening experiments (King and Karaiskos, 2010). Another one

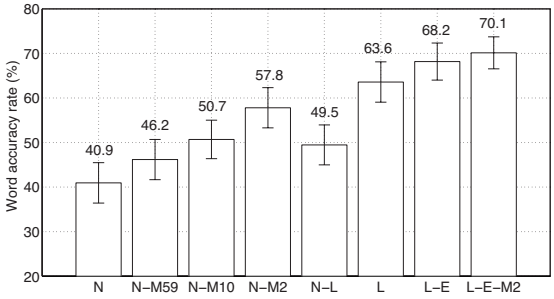


Figure 11: Word accuracy rates for speech-shaped noise (Valentini-Botinhao et al., 2012b).

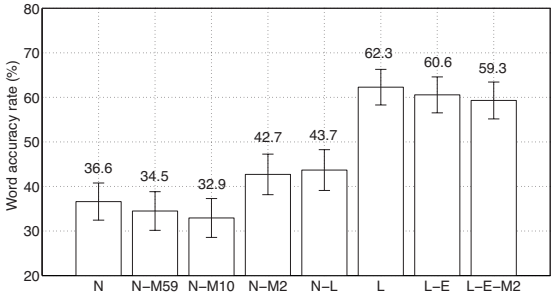


Figure 12: Word accuracy rates for competing talker (Valentini-Botinhao et al., 2012b).

of the sets was used as a practice session done prior to the experiment. All words were considered when calculating the subjective word accuracy rate.

6.5. Results and discussion

We present the mean word accuracy rate (WAR) obtained for each voice when mixed with speech-shaped noise (Figure 11) and a competing talker (Figure 12), along with 95 % confidence intervals.

In the case of ssn, higher intelligibility gains are obtained when we just modified the first few Mel cepstral coefficients (N-M2): coarser frequency modifications were more effective than fine-grained ones (N-M10, N-M59). The best voice generated with modified Mel cepstral coefficients (N-M2) was more intelligible than the Lombard spectral-adapted voice (N-L), and has the advantage of requiring no additional recordings. The fairest comparison system for N-M2 is N-L, since both modify only the spectrum (and not duration or excitation parameters).

The voice which adapts all model parameters to Lombard speech (L) works well, extrapolated adaptation adds a further gain (L-E) and a fur-

ther improvement can be obtained by following this with the proposed Mel cepstral modification (L-E-M2).

As we can see in Figure 12 voices with only spectral modifications (N-M59, N-M10, N-M2 and N-L) obtain only modest gains in the presence of a competing speaker. Higher gains were obtained by the fully adapted Lombard voices (L, L-E, L-E-M2), although the proposed method does not provide additional gains on top of voices L or L-E. This suggests that changes in the spectral envelope contribute less to intelligibility gain than duration or F_0 modifications, for this type of masker noise. Given the non-stationary nature of this masker, we would expect a temporal energy re-allocation strategy (e.g., taking advantage of quiet or silent regions in the noise signal) to be more effective than reallocating energy across different frequencies.

7. Large scale evaluation

Taking the most successful approach based on the proposed method from the last experiment (i.e., the one used to create voice N-M2), we proceeded to evaluate its performance in a larger experiment with more listeners. We compared it to two natural voices – recordings from the normal and Lombard speech dataset – and two other synthetic voices built from normal and Lombard speech. This evaluation was part of the large scale listening experiment described in (Cooke et al., 2013) that used 154 native English speakers. Cooke et al. (2013) also compared several other methods to enhance the intelligibility of natural and synthetic speech in noise. Here we show only the results for our synthetic speech entries, alongside the results for natural normal and Lombard speech.

7.1. Voice building

Using the same natural speech database described in our previous experiment we built three voices for this evaluation. For consistency with (Cooke et al., 2013), we will refer to these voices as TTS, TTSGP and TTSLomb, which correspond to the voices described as N, N-M2 and L above, except for a small difference in the way one voice was used to generate speech. Specifically, we limited the duration modifications induced by voice TTSLomb so that the maximum overall duration increase was no more than half a second per sentence – this was necessary to conform to the rules of the experiment described in (Cooke et al., 2013).

	duration (secs.)	F_0 mean (Hz)	F_0 range (Hz)	spectral tilt (dB/oct.)	loudness (sone)
Natural speech					
Normal	2.06	107.1	34.60	-2.14	11.43
Lombard	2.32	136.8	46.74	-1.83	11.96
Synthetic speech					
TTS	1.95	104.5	22.45	-2.26	10.96
TTSGP				-1.90	12.43
TTSLomb	2.43	145.2	42.55	-1.71	12.06

Table 3: Acoustic properties of the two natural voices (Normal, Lombard) and the three synthetic voices (TTS, TTSGP, TTSLomb).

7.2. Acoustic analysis

We present in Table 3 various acoustic properties calculated per sentence then averaged across the test set: duration changes, prosody changes (in terms of F_0 mean and range), spectral tilt and loudness, calculated using the ISO-532B method (ISO 532, 1975). The F_0 range was calculated as the difference between the 80th and 20th percentiles.

The acoustic changes found here for the natural and synthetic speech data are similar to what has been reported in other studies of Lombard speech data: relative increases in sentence duration (12% natural and 25% synthetic), F_0 mean (27% and 39%) and F_0 range (35% and 90%), flatter spectral tilt (14% and 24%) and increase in loudness (5% and 13%).

The voice TTSGP presents on average a flatter spectral tilt when compared to the TTS voice (16% flatter). TTSGP is slightly louder than the TTSLomb, a relative increase of 13% over the TTS voice. Duration and F_0 remain unchanged, because only the spectral parameters were modified.

7.3. Listening experiment design

To evaluate these voices across a range of SNRs, the five different voices listed in Table 3 were mixed with speech-shaped noise (ssn) and a competing female speaker (cs). The noises were mixed at preselected signal to noise ratios (SNRs) chosen (using a pilot test) to achieve approximately 25, 50 and 75% word accuracy rates in natural unmodified speech (-9 dB, -4 dB, 1 dB for speech-shaped noise and -21 dB, -14 dB, -7 dB for competing talker).

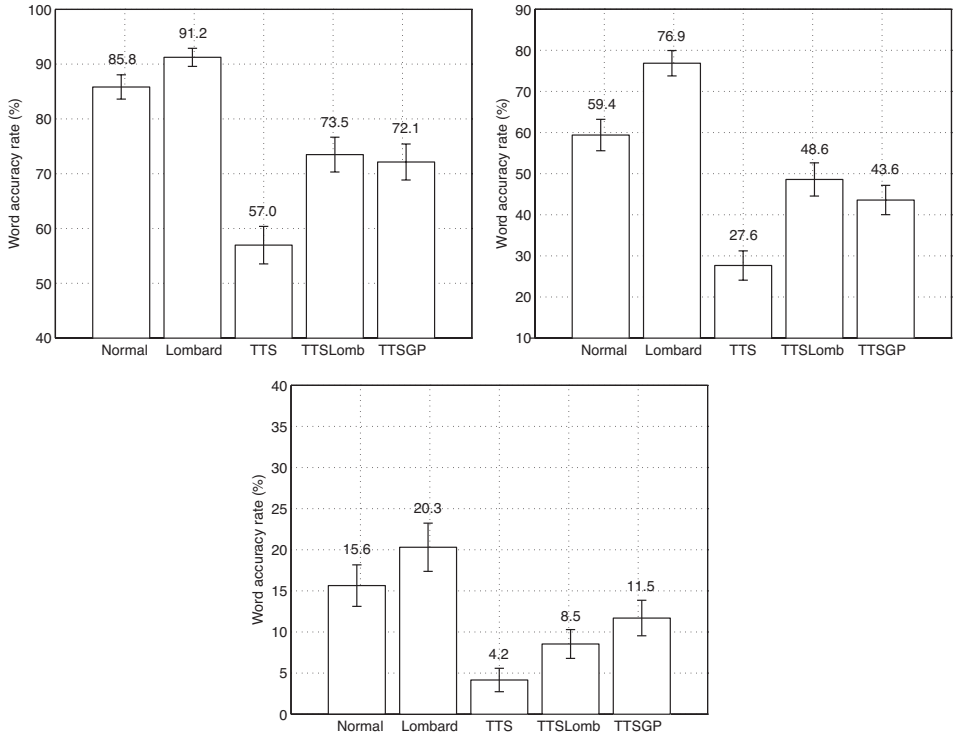


Figure 13: Word accuracy rates for natural voices (Normal, Lombard) and synthetic voices (TTS, TTSLomb, TTSGP) mixed with speech-shaped noise at SNR = 1dB (left), SNR =-4dB (middle), SNR =-9dB (right) (Valentini-Botinhao et al., 2012c).

In total, 154 native English speakers listened to the noisy samples over headphones in sound-isolated booths. 180 sentences from the Harvard corpus were used in a balanced arrangement, such that listeners never heard the same sentence more than once. Each pair of participants between them listened to 5 different sentences of each noise/SNR/voice combination. The subjective word accuracy rates were computed per sentence and – in a procedure improved over the previous listening test – counting only content words (i.e., the words 'a', 'the', 'in', 'to', 'on', 'is', 'and', 'of', 'for' were excluded from scoring).

7.4. Results and discussions

Figures 13 and 14 show the word accuracy rates (WARs) of the five voices mixed with speech-shaped noise and competing speaker for each SNR tested.

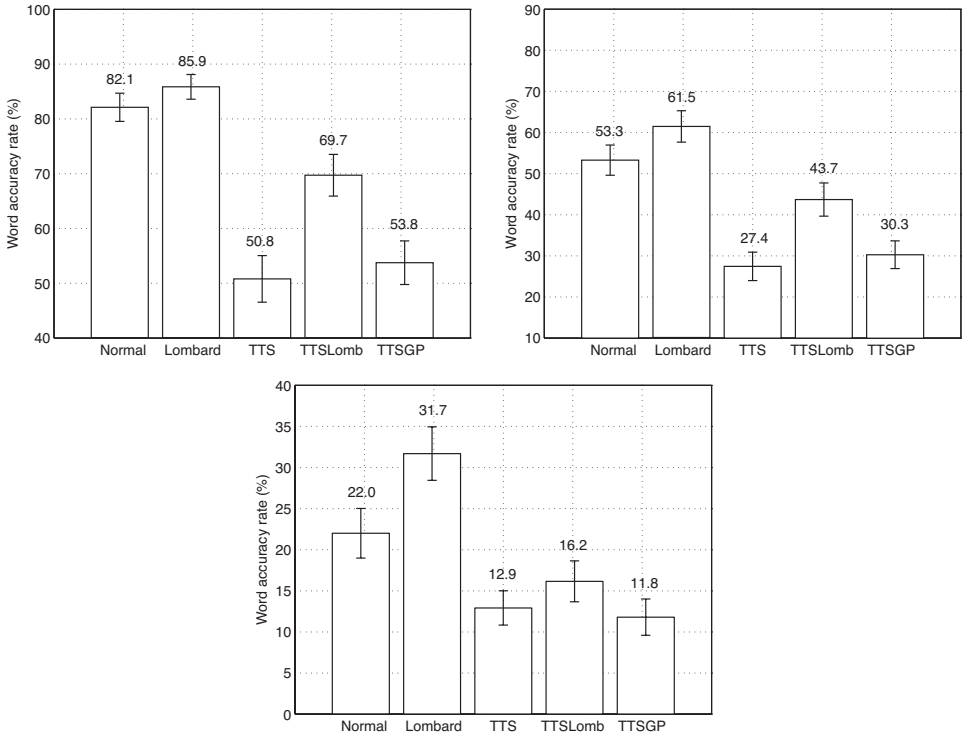


Figure 14: Word accuracy rates for natural voices (Normal, Lombard) and synthetic voices (TTS, TTSLomb, TTSGP) mixed with a female competing speaker at SNR = -7dB (left) SNR =-14dB (middle), SNR =-21dB (right) (Valentini-Botinhao et al., 2012c).

As we can see there is quite a large difference in performance between natural and synthetic voices. In other words, the intelligibility of synthetic speech is much more strongly degraded in the presence of additive noise, compared to natural speech. However, this gap can be closed to a large extent when we modify the spectral envelope using our proposed technique, or when we adapt to Lombard data. The gains obtained when going from natural to Lombard are much larger in the case of the synthetic speech than for the natural speech (average across SNRs: 47% vs. 17% for ssn; 42% vs. 13% for cs).

The proposed method, which does not required re-training of the synthesis models, created a voice (TTSGP) that provides intelligibility gains over a normal synthetic voice (TTS). The word accuracy rates obtained by the TTSGP voice are comparable to those obtained with the TTSLomb voice, in the case of speech-shaped noise, even though the proposed method does

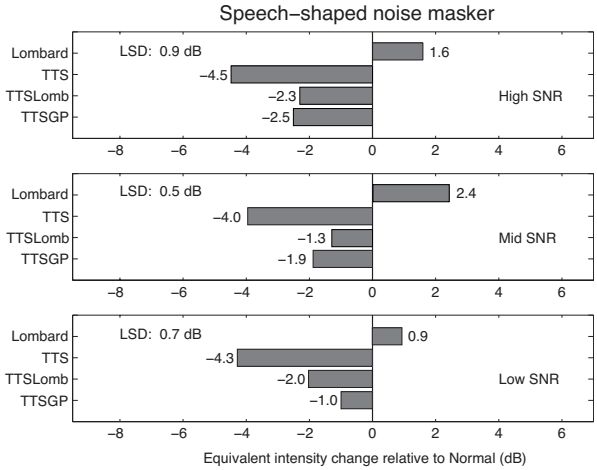


Figure 15: Equivalent intensity change relative to natural speech, for the case of speech-shaped noise. LSD indicates Fisher's least significant different converted to dB via the psychometric function for this masker. Adapted from (Cooke et al., 2013).

not modify duration or F_0, and requires no additional data. Averaged across SNRs, the relative gains obtained by TTSGP over the TTS voice were 44% for ssn and 5% for cs. Once again, we see that in the presence of a competing talker masker, only moderate improvements are observed, suggesting the greater importance of prosody and duration in this condition.

The method employed in (Cooke et al., 2013) uses a psychometric function which means we are able to express the change in intelligibility in terms of "equivalent intensity change" relative to normal natural speech, which is an intuitively appealing way of presenting the results on a dB scale. This is shown in Figures 15 and 16 for ssn and cs. We can see the effective loss (in dB) of using synthetic speech compared to natural speech (average across SNR: TTS $-4.3\,$dB for ssn and $-5.9\,$dB for cs) and how this loss can be substantially mitigated by we modifying the synthetic voice spectral envelope using our proposed method (TTSGP $-1.8\,$dB for ssn and $-5.6\,$dB for cs) or by adapting the models to Lombard speech from the same speaker (TTSLomb $-1.9\,$dB for ssn and $-2.7\,$dB for cs).

8. Conclusions

We have presented a method for increasing the intelligibility of HMM-generated synthetic speech in the presence of noise, based on the glimpse

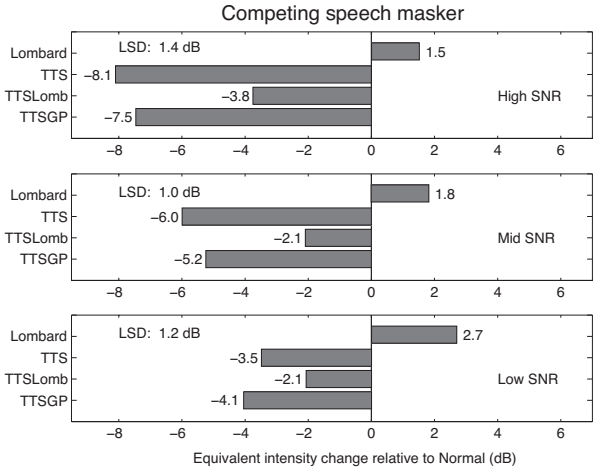


Figure 16: Equivalent intensity change relative to natural speech, for the case of competing speaker noise. LSD indicates Fisher's least significant different converted to dB via the psychometric function for this masker. Adapted from (Cooke et al., 2013).

proportion measure. The method operates on the Mel cepstral coefficients generated by acoustic models which have been previously trained only on natural read speech collected in quiet conditions, of the type normally used to build text-to-speech systems. The method updates the Mel cepstral coefficients iteratively via gradient descent such that the glimpse proportion increases, without changing the overall energy. We have observed that sentences generated with such modified Mel cepstral coefficients have a boost in frequencies between 1-4 kHz and that this boost is highly dependent on the phonetic units (vowels and nasals are more more enhanced than fricatives and stops). Results with a speech-shaped noise masker show that the modified voice is as intelligible as a synthetic voice trained with plain speech then adapted to Lombard speech. When mixed with a competing talker the gains are more modest for both the proposed method and for adaptation to Lombard speech.

9. Future work

While speech intelligibility increases, naturalness and quality might have been compromised, especially if the modified speech is heard in clean conditions. To decrease the artefacts that could arise from the frame-by-frame processing we have a few ideas. One would be to apply the optimization

method to the static components of the spectral statistical models instead so the maximum likelihood parameter generation (Tokuda et al., 2000) could smooth the differences between the consecutive frames of modified spectral parameters. We would also like to investigate whether updating the spectral coefficients at a slower analysis window rate can decrease artefacts while still maintaining intelligibility gains. To evaluate this , we plan to perform preference listening tests in clean and noisy conditions as well as task oriented experiments that would allow a longer exposure to modified speech. Ongoing and future work also includes a more extensive comparison with a wider variety of other intelligibility enhancement methods, and investigation of methods that can reallocate energy across time.

10. Acknowledgement

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213850 (SCALE) and 256230 (LISTA), and from EPSRC grants EP/I031022/1 (NST) and EP/J002526/1 (CAF).

References

B. Picart, T. Drugman, T.D., 2011. Continuous control of the degree of articulation in hmm based speech synthesis, in: Proc. Interspeech, Florence, Italy.

Bonardo, D., Zovato, E., 2007. Speech synthesis enhancement in noisy environments, in: Proc. Interspeech, Antwerp, Belgium. pp. 2853–2856.

Castellanos, A., Benedi, J., Casacuberta, F., 1996. An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. Speech Comm. 20, 23 – 35.

Cooke, M., 2003. Glimpsing speech. Journal of Phonetics 31, 579 – 584.

Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119, 1562–1573.

Cooke, M., Lu, Y., 2010. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. J. Acoust. Soc. Am. 128, 2059–2069.

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 Comm. .

Dreschler, W., Verschuure, H., Ludvigsen, C., Westermann, S., 2001. ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology. Audiology 40, 148–57.

Fukada, T., Tokuda, K., Kobayashi, T., Imai, S., 1992. An adaptive algorithm for mel-cepstral analysis of speech, in: Proc. ICASSP, San Francisco, USA. pp. 137–140.

Garnier, M., Bailly, L., Dohen, M., Welby, P., Loevenbruck, H., 2006. An acoustic and articulatory study of Lombard speech: global effects on the utterance, in: Proc. ICSLP, Pittsburgh, USA. pp. 2246–2249.

Hansen, J., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Comm. 20, 151 – 173.

Howell, P., Barry, W., Vinson, D., 2006. Strength of British English accents in altered listening conditions. Perception and Psychophysics 68, 139–153.

IEEE, 1969. IEEE recommnded pratice for speech quality measurements. IEEE Trans. on Audio and Electroacoustics 17, 225 – 246.

ISO 532, 1975. Acoustics - method for calculating loudness level.

Junqua, J., 1993. The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. 93, 510–524.

Kawahara, H., Masuda-Katsuse, I., Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Comm. 27, 187–207.

King, S., Karaiskos, V., 2010. The Blizzard Challenge 2010, in: Proc. Blizzard Challenge Workshop, Kyoto, Japan.

Koishida, K., Tokuda, K., Kobayashi, T., Imai, S., 1996. CELP coding system based on mel-generalized cepstral analysis, in: Proc. ICSLP, pp. 318 –321.

Koishida, K., Tokuda, K., Kobayashi, T., Imai, S., 2000. Spectral representation of speech based on mel-generalized cepstral coefficients and its properties. Electronics and Communications in Japan 83, 50–59.

Langner, B., Black, A.W., 2005. Improving the understandability of speech synthesis by modeling speech in noise, in: Proc. ICASSP, pp. 265–268.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H&H theory, in: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publishers, pp. 403–439.

Lu, Y., Cooke, M., 2008. Speech production modifications produced by competing talkers, babble, and stationary noise. J. Acoust. Soc. Am. 124, 3261–3275.

McLoughlin, I., Chance, R., 1997. LSP-based speech modification for intelligibility enhancement, in: Proc. Digital Signal Processing, Santorini , Greece. pp. 591–594.

Moore, B.C.J., Glasberg, B.R., 1996. A revision of Zwicker's loudness model. Acta Acustica 82, 335–345.

Nicolao, M., Latorre, J., Moore, R.K., 2012. C2H A computational model of H&H-based phonetic contrast in synthetic speech, in: Proc. Interspeech, Portland, USA.

Patel, R., Schell, K.W., 2008. The influence of linguistic content on the Lombard effect. J. Speech Lang. Hear. Res. 51, 209–220.

Picheny, M., Durlach, N., Braida, L., 1985. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. J. Speech Hear. Res. 28, 96–103.

Raitio, T., Suni, A., Vainio, M., Alku, P., 2011. Analysis of HMM-based lombard speech synthesis, in: Proc. Interspeech, Florence, Italy.

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

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

Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., Stokes, M., 1988. Effects of noise on speech production: Acoustic and perceptual analysis. J. Acoust. Soc. Am. 84, 917–928.

Suni, A., Raitio, T., Vainio, M., Alku, P., 2010. The GlottHMM speech synthesis entry for Blizzard Challenge 2010, in: Proc. Blizzard Challenge Workshop, Kyoto, Japan.

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

Tang, Y., Cooke, M., 2010. Energy reallocation strategies for speech enhancement in known noise conditions, in: Proc. Interspeech, pp. 1636–1639.

Tang, Y., Cooke, M., 2012. Optimised spectral weightings for noise-dependent speech intelligibility enhancement, in: Proc. Interspeech, Portland, USA.

Tang, Y., Cooke, M., Valentini-Botinhao, C., 2013. A distortion-weighted glimpse-based intelligibility metric for modied and synthetic speech, in: Proc. SPIN.

Toda, T., Tokuda, K., 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans. Inf. Syst. E90-D, 816–824.

Tokuda, K., Kobayashi, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis, in: Proc. ICASSP, pp. 1315–1318.

Valentini-Botinhao, C., Maia, R., Yamagishi, J., King, S., Zen, H., 2012a. Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise, in: Proc. ICASSP, Kyoto, Japan.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2011a. Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?, in: Proc. Interspeech, Florence, Italy.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2011b. Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise, in: Proc. ICASSP, Prague, Czech Republic.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2012b. Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise, in: Proc. Interspeech, Portland, USA.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2012c. Speech intelligibility enhancement for HMM-based synthetic speech in noise, in: Proc. SAPA, Portland, USA.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. taslp 17, 66 –83.

Yamagishi, J., Zen, H., Wu, Y.J., Toda, T., Tokuda, K., 2008. Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge, in: Proc. Blizzard Challenge Workshop, Brisbane, Australia.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Comm. 51, 1039–1064.