Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise

Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King

The Centre for Speech Technology Research, University of Edinburgh, UK

C.Valentini-Botinhao@sms.ed.ac.uk, jyamagis@inf.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

We propose a method that modifies the Mel cepstral coefficients of HMM-generated synthetic speech in order to increase the intelligibility of the generated speech when heard by a listener in the presence of a known noise. This method is based on an approximation we previously proposed for the Glimpse Proportion measure. Here we show how to update the Mel cepstral coefficients using this measure as an optimization criterion and how to control the amount of distortion by limiting the frequency resolution of the modifications. To evaluate the method we built eight different voices from normal read-text speech data from a male speaker. Some voices were also built from Lombard speech data produced by the same speaker. Listening experiments with speech-shaped noise and with a single competing talker indicate that our method significantly improves intelligibility when compared to unmodified synthetic speech. The voices built from Lombard speech outperformed the proposed method particularly for the competing talker case. However, compared to a voice using only the spectral parameters from Lombard speech, the proposed method obtains similar or higher performance.

Index Terms: intelligibility of speech in noise, Mel cepstral coefficients, HMM-based speech synthesis

1. Introduction

Humans change their speaking style when conversing in a noisy environment so that communication success is ensured, often producing what is called Lombard speech. It is unclear what aspects of Lombard speech actually contribute to intelligibility increases and how they relate to the nature of the noise. Solving this problem will enable practical applications which automatically modify natural or synthetic speech to increase intelligibility in noise.

The parametrical statistical framework of HMM-based speech synthesis offers many different ways to approach this problem. If Lombard speech data are available for the speaker whose TTS voice we want to modify, we can use adaptation techniques to produce new Lombard-like speech for that speaker [1]. If such data are not available, then we can apply noise-independent modifications at the feature level based on known acoustic properties of Lombard speech, such as F0 increase, flattening of spectral tilt and duration stretch [1]. However if we want to employ noise-dependent techniques then we need to be able to automatically detect what sort of modifications should take place for certain pairs of speech and noise signals. One way in which this can be done is by using an intelligibility measure of speech [2]. Such an approach is limited by the performance of the objective measure: if it fails to accurately predict intelligibility then any modification based on that prediction is likely to fail. Therefore, it is important to find a specific domain of modifications where the intelligibility model behaves well and ensure that the modifications applied in this domain remain within the working range of the objective model.

We have observed that the Glimpse Proportion (GP) measure for speech intelligibility in noise [3] has a high correlation coefficient with subjective intelligibility scores for HMM-generated synthetic speech whose spectral envelope has been modified [4]. Moreover, modifications in the spectral envelope domain can achieve quite high intelligibility gains. We then proposed a cepstral extraction method based on the GP measure for the HMM-based synthesis framework [5]. This method was shown to provide significant intelligibility improvement, although not for all noise types. We hypothesise this is due to distortions introduced by the method itself. A disadvantage of that approach is having to train a different model for each noise type, because the noise-dependent modifications are performed as part of feature extraction. Now, we propose a method that can be applied at generation time, and not requiring any information about the spectral envelope of natural speech to achieve distortion control. Rather, we propose to control the distortion in two ways: using a stopping criteria based on the mismatch between the auditory representations of modified and unmodified speech, as proposed by the GP measure, and only modifying the first few cepstral coefficients, thus limiting the frequency resolution of the modifications. A further extension proposed in this paper is the possibility of using this method for *Mel* cepstral coefficients, which can provide higher speech quality with fewer coefficients [6].

In Section 2 and 3 we show how Mel cepstral coefficients model the spectrum, how the GP measure works and how we previously approximated it for the purpose of cepstral coefficient optimization. In Section 4 we introduce the new method for Mel cepstral modification based on the GP measure. We then provide experimental results from listening experiments to support our conclusions.

2. Mel cepstral coefficients

We can represent the spectrum by M-th order Mel cepstral coefficients $\{c_m\}_{m=0}^{M}$ in the following manner [6]:

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

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

where α is a warping factor which can be chosen to represent, for instance, the Mel scale [6].

3. The Glimpse Proportion measure

The Glimpse Proportion (GP) measure for speech intelligibility in noise [3] is the proportion of spectral-temporal regions called glimpses where speech is more energetic than noise. The motivation behind this measure is that when humans listen to speech in noise they tend to focus on such regions. The Spectro Temporal Excitation Pattern (STEP) representation used by the measure is obtained in the following manner: Gammatone filtering, envelope extraction and smoothing, averaging over a time frame and level compression [3].

In [5] we showed how to approximate the GP measure in a way that provides a closed and differentiable formulation:

$$GP = \frac{100}{N_f N_t} \sum_{t=1}^{N_t} \sum_{f=1}^{N_f} \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \tag{3}$$

where N_t and N_f are the number of time frames and frequency channels, $\mathcal{L}(.)$ is a logistic sigmoid function of zero offset and slope η, $y_{t,f}^{sp}$ and $y_{t,f}^{ns}$ are the approximated STEP representations for speech and noise respectively at analysis window t and frequency channel f.

The STEP representation for speech is given by:

$$y_{t,f}^{sp} = \frac{1}{N} (\boldsymbol{G}_f \boldsymbol{h}_t \circledN \boldsymbol{G}_f \boldsymbol{h}_t)^\top \boldsymbol{S}\, \boldsymbol{b} \tag{4}$$

where N is the number of frequency bins of the spectrum, $\circledN$ is circular convolution of dimension N, $\boldsymbol{h}_t$ is an Nx1 vector containing the magnitude spectrum of windowed speech signal at analysis window t, $\boldsymbol{G}_f$ is an NxN diagonal matrix whose diagonal contains the Gammatone filter frequency response for frequency channel f, $\boldsymbol{S}$ is an NxN diagonal matrix whose diagonal contains the frequency response of the smoothing filter and $\boldsymbol{b}$ is an Nx1 vector containing the coefficients of the average filter.

4. Mel cepstral modifications based on the GP measure

Given a set of Mel cepstral coefficients and a noise signal we want to obtain a new set of Mel cepstral coefficients $\boldsymbol{c}_t = [c_{t,1} \ \ldots \ c_{t,m} \ \ldots \ c_{t,M}]^\top$ that maximizes GP_t, the value of the function described in Eq. (3) in time frame t. We then have:

$$\boldsymbol{c}_t = \operatorname{argmax} GP_t \tag{5}$$

$$GP_t = \frac{100}{N_f} \sum_{f=1}^{N_f} \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \tag{6}$$

As this function is not necessarily convex with respect to the Mel cepstral coefficients, we use a Steepest Descent method to solve the optimization. The update equation is:

$$\boldsymbol{c}_t^{(i+1)} = \boldsymbol{c}_t^{(i)} + \mu \boldsymbol{\nabla} GP_t^{(i)} \tag{7}$$

where μ is the step size and the i index refers to iterations. From now on we drop the i index for clarity. The gradient vector is given by:

$$\boldsymbol{\nabla} GP_t = \frac{100}{N_f N} \sum_{f=1}^{N_f} \eta\, \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \big[1 - \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})\big] \cdot \boldsymbol{H}_{c_t} \boldsymbol{G}_f \left(2\, \boldsymbol{\Gamma}_N \circledN \boldsymbol{G}_f \boldsymbol{h}_t\right) \boldsymbol{S}\, \boldsymbol{b} \tag{8}$$

where $\boldsymbol{H}_{c_t}$ is an MxN matrix whose elements are $\{\boldsymbol{H}_{c_t}\}_{m,j} = \frac{\partial |H_t(\omega_j)|}{\partial c_{t,m}}$ and the operation $(\boldsymbol{\Gamma}_N \circledN \boldsymbol{G}_f \boldsymbol{h}_t)$ defines an NxN matrix whose n-th row is equal to $\boldsymbol{e}_n \circledN (\boldsymbol{G}_f \boldsymbol{h}_t)^\top$, $\boldsymbol{e}_n$ being the n-th column of the identity matrix $\boldsymbol{\Gamma}_N$.

When the spectrum is modelled by Mel cepstral coefficients as defined in Eq.(1) the elements of the matrix $\boldsymbol{H}_{c_t}$ are given by:

$$\frac{\partial |H_t(\omega_j)|}{\partial c_{t,m}} = |H_t(\omega_j)| \cos(m\, \tilde{\omega}_j) \tag{9}$$

However because we do not wish to modify the energy of the speech signal we have:

$$\frac{\partial |H_t'(\omega_j)|}{\partial c_{t,m}} = |H_t'(\omega_j)| \Big(\cos(m\, \tilde{\omega}_j) - \frac{1}{\psi} \sum_{l=1}^{N} |H_t(\omega_l)|^2 \cos(m\, \tilde{\omega}_l) \Big) \tag{10}$$

where $|H_t'(\omega_j)|$ is the energy-normalized magnitude spectrum and $\psi = \sum_{j=1}^{N} |H_t(\omega_j)|^2$. There is no need to update the first Mel cepstral coefficient c_0 as the normalization operation updates it to a certain value regardless of an additional Δc_0 term.

An issue we face when using the GP measure as an optimization criterion on its own is the need to limit the distortions caused by the modifications. 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 whilst at the same time limiting the frequency resolution of the modifications. To implement that, we simply set the gradient vector for higher dimensions to zero, thus modify only the first few Mel cepstral coefficients, which represent the coarse properties of the spectrum.

5. Evaluation

In this section we show how we built the TTS voices, give an acoustic analysis, and present the results of a listening test.

5.1. Voice building

To build the voices used in this evaluation we used two different datasets recorded by the same British male speaker: normal (plain, read-text) speech data and Lombard speech. The Lombard dataset was recorded while the speaker listened to speech-modulated noise based on another male speaker [7] played over headphones at a absolute value of 84 dBA.

We built eight different voices as outlined in Table 1. Voice N was created from a high quality average voice model adapted to 2803 sentences of the normal speech database, corresponding to three hours of material. We decided to use an average voice model rather than building a speaker-dependent voice because the normal speech dataset was not phonetically balanced. Voices N-M59, N-M10 and N-M2 are variations of N in which we modify all, just the first ten (c_1 until c_{10}), or just the first two (c_1 and c_2) Mel cepstral coefficients using our proposed method.

Lombard voice L was based on voice N, further adapted using 780 sentences from the Lombard speech dataset, corresponding to 53 minutes of recorded material. Again, the reason for using adaptation was the lack of phonetic balance in the speech dataset. Voice N-L was also created from voice N but

Voice	Adaptation	Modification
N	-	-
N-M59	-	all coefficients
N-M10	-	first 10 coefficients
N-M2	-	first 2 coefficients
N-L	only spectral parameters	-
L	all dimensions	-
L-E	all dimensions extrapolated	-
L-E-M2	all dimensions extrapolated	first 2 coefficients

Table 1: *Voices built for the evaluation.*

this time only the Mel cepstral coefficients were adapted to the Lombard data. Voices L-E and L-E-M2 are versions of voice L where we extrapolated the adaptation (voice L-E), and then also modified the two first Mel cepstral using the proposed method (voice L-E-M2).

The training and adaptation data had a sampling rate of 48 kHz. To train, adapt and generate speech we extracted: 59 Mel cepstral coefficients with $\alpha = 0.77$, Mel scale F0, and 25 aperiodicity energy bands extracted using STRAIGHT [8]. We used a hidden semi-Markov model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values; one stream for the spectrum, three streams for the logF0 and one for the band-limited aperiodicity. The Global Variance method [9] was also applied to compensate for the over smoothing effect of the acoustical modelling.

To modify the generated Mel cepstral coefficients we used the method proposed in the previous section, obtaining the STEP representation by using Gammatone filters that covered the range 50-7500 Hz as the noise signal used for testing is sampled at 16 kHz. The stepsize was normalized: $\mu^{(i)} = \mu / ||\nabla GP_t^{(i)}||$ and we set $\mu = 0.4$ for N-M59 and $\mu = 0.8$ for N-M10 and N-M2. We used as stopping criteria both error convergence and a maximum distortion threshold set to be 10 % of relative increase in the Euclidian distance between the STEP representation of original and modified speech.

5.2. Acoustic analysis

Fig.1 shows the Long Term Average Spectrum (LTAS) of the normal (N), modified (N-M2) and Lombard (L) voices, for the case of speech-shaped noise. Compared to voice N, voice N-M2 exhibits enhanced energy in the frequency region of 1-4 kHz and attenuated below 1 kHz. Voice L shows enhancement and attenuation in the same regions as N-M2, although these changes are not as pronounced, attenuation is also seen between 4-5.5 kHz and enhancement at frequencies above this.

Table 2 provides an acoustic analysis of the voices – average duration of speech and pauses, average spectral tilt, and F0 – across all sentences used in the listening test for the normal (N), modified (N-M2) and lombard (L) voices. We can see that, as expected, the Lombard voice produces sentences with longer duration and longer pauses, greatly increased F0 mean and flattening of the spectral tilt. The spectral tilt reflects changes in both spectral envelope and excitation signal. The modified voice N-M2 also presents a flatter spectral tilt, though not to the same extent as the Lombard voice.

5.3. Listening experiments design

We mixed the eight different synthetic voices with two noises: speech-shaped noise and speech from a single competing female talker. For intelligibility testing, it is important to avoid floor or ceiling effects on word error rate. Therefore, in order to obtain intelligibility scores in similar ranges for each noise, we mixed them at differing SNRs: -4 dB for speech-shaped noise and -14 dB for the competing talker. Across the different voices we made sure that the root mean square value was the same.

For the listening test we used 32 native English speakers listening to the noisy samples over headphones in soundproof booths and typing in what he or she heard. Each participant heard six different sentences per condition, i.e., voice and noise type, and each sentence could only be played once. We used the first ten sets of the Harvard sentences [10]; another one of the sets was used as a practice session which listeners completed before the test proper.

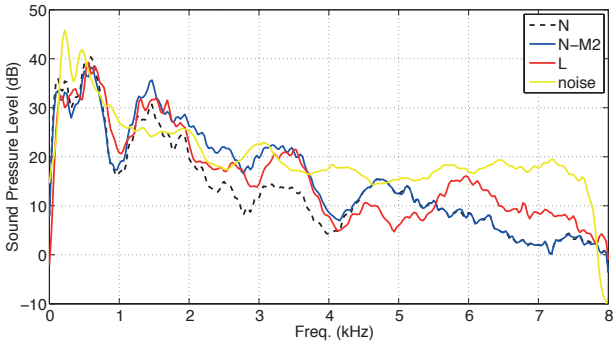

Figure 1: *Long term average spectrum of the normal N, normal modified N-M2 and Lombard L voices for speech-shaped noise.*

Voice	speech (secs.)	pauses (secs.)	F0 (Hz)	spectral tilt (dB/octave)
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.*

5.4. Results and discussion

Figs. 2 and 3 show the mean word accuracy rate (WAR) obtained by each voice when mixed with speech-shaped noise and a competing talker respectively, along with 95 % confidence intervals. Fig.2 shows that the modified voices N-M59, N-M10 and N-M2 achieve higher WAR than the unmodified voices N (40.9 %), and this is significantly higher for the N-M10 (50.6 %) and N-M2 (57.8 %). The N-M2 voice obtains a higher WAR than the N-L voice (49.4 %). The Lombard voices L (63.5 %), L-E (68.1 %) and L-E-M2 (70.1 %) performed better than the normal speech voices although we did not find a significant difference between N-M2 and L. The extrapolated voice L-E is more intelligible than voice L, a trend that is further enhanced by applying our modifications to it, as in voice L-E-M2. The results obtained for the competing talker situation are displayed in Fig. 3 and show a slightly different trend. There is a drop in performance for N-M59 and N-M10 when compared to N (36.6 %), although this is not significant. The N-M2 (42.7 %) voice performs better than the unmodified counterpart N and obtains a similar WAR to N-L (43.6 %). All Lombard voices performed significantly better than the other voices, in particu-

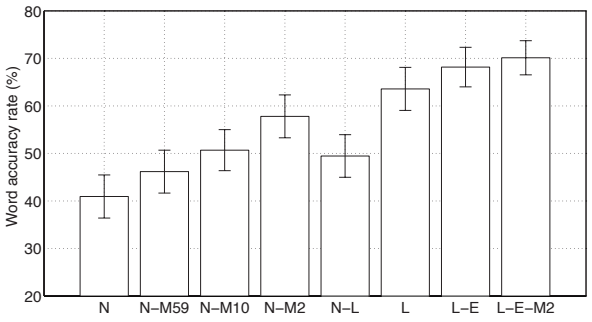

Figure 2: *Word accuracy rates for speech-shaped noise.*

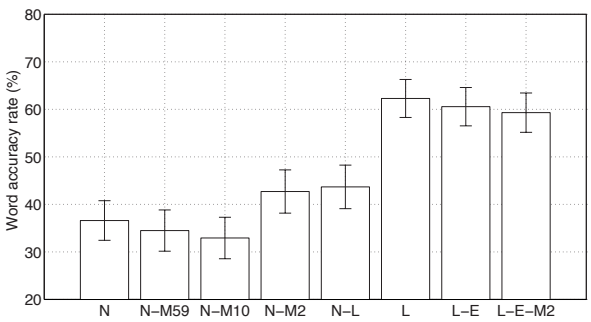

Figure 3: *Word accuracy rates for competing talker.*

lar the L voice (62.2 %). The other versions, L-E (60.5 %) and L-E-M2 (59.3 %), do not appear to increase intelligibility.

As predicted by our hypothesis that distortions were defeating potential gains in intelligibility in our previously-published experiments [5], the voices where we modify only the first few Mel cepstral coefficients achieved a better WAR, indicating that very fine frequency modifications cause distortions that cancel out any potential intelligibility gain they may offer. Compared to the N-L voice, for which the spectral parameters were obtained using Lombard speech, the modifications proposed here obtained a similar or higher intelligibility score. The intelligibility gains obtained by the full Lombard voice L over the N-L voice reflect the impact of changes in duration patterns, F0 and the aperiodicity parameters that define the excitation signal, as pointed out in Table 2. We can see, then, that there is a lot to gain from modifying those parameters in addition to the spectral ones. The spectral modifications proposed here increased the gains obtained with the Lombard voice for speech-shaped noise, as we can see from the results for voice L-E-M2, which shows that there are still gains to be had over and above simply building voices on recorded Lombard speech.

For the competing talker, spectral changes seem to contribute less than for speech-shaped noise. For the competing talker, duration stretches as well as F0 increases are more important. This suggests that for non-stationary noise it is more effective to perform temporal energy re-allocation (e.g., taking advantage of quiet or silent regions in the noise signal) than it is to reallocate energy across different frequencies.

6. Conclusions

We have proposed a new method for modifying Mel cepstral coefficients based on an intelligibility measure for speech in noise, the Glimpse proportion measure. We showed how to control distortion by modifying only the first few Mel cepstral coefficients, which is a natural way of limiting the frequency resolution of the modifications. In the evaluation, we compared synthetic voices whose spectral parameters were modified as well as using spectral parameters from Lombard speech. Listening tests using speech-shaped noise and a competing talker indicate that we only need to modify two Mel cepstral coefficients to obtain a similar or higher intelligibility to Lombard spectral modifications. Moreover we observed that, for the competing talker, the intelligibility gain obtained by the Lombard voice over the modified voice was mainly due to changes in duration, F0 and excitation parameters. In terms of what can be achieved when modifying only Mel cepstral coefficients, our method obtains either higher or similar intelligibility scores to Lombard Mel cepstral coefficients. We are currently making a more extensive comparison of our method to other intelligibility enhancement methods. In future, we plan to investigate reallocating energy across time. We also plan operating under a loudness constraint rather than an energy one.

7. Acknowledgment

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 and EP/J002526/1.

8. References

[1] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM-based lombard speech synthesis," in *Proc. Interspeech*, Florence, Italy, August 2011.

[2] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. ICASSP*, Toulouse, France, May 2006, p. 493496.

[3] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[4] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. Interspeech*, Florence, Italy, August 2011.

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

[6] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, vol. 1, San Francisco, USA, March 1992, pp. 137–140.

[7] W. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology." *Audiology*, vol. 40, no. 3, pp. 148–57, 2001.

[8] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "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.*, vol. 27, pp. 187–207, 1999.

[9] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[10] "IEEE recommended pratice for speech quality measurements," *Audio and Electroacoustics, IEEE Transactions on*, vol. 17, no. 3, pp. 225 – 246, sep 1969.