CEPSTRAL ANALYSIS BASED ON THE GLIMPSE PROPORTION MEASURE FOR IMPROVING THE INTELLIGIBILITY OF HMM-BASED SYNTHETIC SPEECH IN NOISE

Cassia Valentini-Botinhao[1]*, Ranniery Maia*[2]*, Junichi Yamagishi*[1]*, Simon King*[1] *and Heiga Zen*[2]

[1] The Centre for Speech Technology Research, University of Edinburgh, UK
[2] Cambridge Research Laboratory, Toshiba Research Europe Limited, UK

ABSTRACT

In this paper we introduce a new cepstral coefficient extraction method based on an intelligibility measure for speech in noise, the Glimpse Proportion measure. This new method aims to increase the intelligibility of speech in noise by modifying the clean speech, and has applications in scenarios such as public announcement and car navigation systems. We first explain how the Glimpse Proportion measure operates and further show how we approximated it to integrate it into an existing spectral envelope parameter extraction method commonly used in the HMM-based speech synthesis framework. We then demonstrate how this new method changes the modelled spectrum according to the characteristics of the noise and show results for a listening test with vocoded and HMM-based synthetic speech. The test indicates that the proposed method can significantly improve intelligibility of synthetic speech in speech shaped noise.

Index Terms— cepstral coefficient extraction, objective measure for speech intelligibility, Lombard speech, HMM-based speech synthesis

1. INTRODUCTION

This work focuses on compensating for background additive noise by increasing the intelligibility of synthetic speech generated by a parametric statistical model. Our method modifies clean speech before it is added to noise. Applications of such an approach include car navigation systems and any public announcement system that makes use of text to speech technology.

Intelligibility of state-of-the-art hidden Markov model (HMM) generated synthetic speech can be comparable to natural speech in clean environments [1] but in noisy environments the situation is quite different and most often natural speech is more intelligible. The statistical and parametric nature of HMM-based speech synthesis however offers a high degree of control over the generated speech. By modifying the models or extracted parameters we are able to control the acoustic characteristics of the generated speech without the need for new data. It is then possible to generate synthetic speech that is more intelligible in noise than the natural speech used for training [2]. One way to achieve this is to imitate the acoustic properties found in natural speech produced in noise, also known as Lombard speech. However not all observed acoustic changes improve intelligibility. It has for example been found that changes in the fundamental frequency have little contribution to intelligibility gains [3, 4]. What remains unknown is which acoustic modifications do in fact have a positive impact on intelligibility and how they relate to the noise characteristics.

We believe that it is possible to increase the intelligibility of speech in noise by modifying clean speech automatically according to the noise characteristics. Because we do not know how speech production and background noise are related, we need a model of intelligibility, or just simply an objective measure for speech intelligibility in noise, to control how speech should be modified. This is what we refer here as an auditory perceptual based approach, as the modifications are no longer inspired by speech production in noise but by how the human auditory system perceives them. Previously we have shown that simple changes in the spectral domain can result in significant gains in intelligibility for HMM-generated synthetic speech in noise and that some intelligibility measures can predict these intelligibility gains [4]. Our idea here is then to use one of these measures, the Glimpse Proportion (GP) measure [5], to modify the spectral envelope of speech. To do this we alter the optimization criterion of the cepstral coefficient extraction method [6] commonly used in the HMM-based synthesis framework.

In Section 2 of this paper we outline the cepstral coefficient extraction method and in Section 3 we describe the Glimpse Proportion measure. In Section 4 we show how we can reformulate the Glimpse Proportion measure to use as a cost function for cepstral extraction and then we define the proposed cepstral extraction method, showing how to solve the new optimization problem. Section 5 gives the experimental results on the acoustic analysis of the modifications and intelligibility evaluation of vocoded and HMM-generated synthetic speech.

2. UELS-BASED CEPSTRAL ANALYSIS

The cepstral coefficient extraction as described in [6] is a method commonly used to extract spectral parameters for an HMM-based speech synthesizer. The method is based on the Unbiased Estimator of Log Spectrum (UELS) [7].

The cepstral coefficients $\{c_m\}_{m=0}^{M}$ define the spectrum of the speech signal $s(n)$ in the following way:

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

where $K = \exp c_0$ and $D(e^{j\omega})$ is the gain normalized version of $H(e^{j\omega})$.

The authors in [6] propose to extract cepstral coefficients by minimizing the criterion defined for the unbiased condition as described in [7]. Since $H(e^{j\omega})$ as defined in Eq. (1) is a minimum phase system it is possible to prove that minimizing the unbiased criterion with respect to $\{c_m\}_{m=1}^{M}$ is the same as minimizing the following cost function:

$$\varepsilon = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_N(\omega)}{|D(e^{j\omega})|^2}\, d\omega \tag{2}$$

where $I_N(\omega)$ is the modified periodogram of a wide-sense stationary process $s(n)$. Likewise we find that $K = \sqrt{\varepsilon_{min}}$, where ε_{min} is the minimum value of ε.

3. THE GLIMPSE PROPORTION MEASURE

The Glimpse Proportion (GP) measure for speech intelligibility in noise [5] is based on the idea that in a noisy environment humans focus on glimpses of speech that are not masked by noise. It correlates well with subjective scores for intelligibility in noise of both natural [5] and HMM-based synthetic speech [8] and also when the spectral envelope of HMM-based synthetic speech is modified [4]. The GP measure outperforms most existing measures for intelligibility of speech in noise and it does not require any time delays.

The measure is the proportion of spectral-temporal regions, so called glimpses, where speech is more energetic than noise. This comparison takes place in the Spectro Temporal Excitation Pattern (STEP). In order to represent a signal in this domain the following operations are performed over the speech and noise waveform separately: Gammatone filtering into frequency channel, envelope extraction, envelope smoothing, average over time frame and level compression. The centre frequencies of the Gammatone filters are linearly spaced in the equivalent rectangular bandwidth (ERB) scale [9].

4. PROPOSED CESPTRAL ANALYSIS INCORPORATING THE GP MEASURE

In this section we show how we can approximate the GP measure and integrate it to the existing cost function for cepstral coefficient extraction shown in Section 2.

4.1. Proposed GP approximation

To obtain a closed and differentiable formula that relates spectral parameters to the Glimpse Proportion measure we have to make some approximations and correspondences. We first replace the hard decision for counting glimpses by a soft one defined by a sigmoid function. The proposed approximated Glimpse Proportion measure is then given by:

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

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

We approximate the calculation of the STEP signal for speech by performing it over the magnitude spectrum of speech. The absolute value operation representing the envelope extraction step is replaced by a circular convolution of the signal with itself. The filtering operations are replaced by truncated multiplications and the level compression is no longer considered. The STEP approximation for the noise signal differs only in the fact that the input signal is not the magnitude spectrum but the discrete Fourier transform representation of the noise. The STEP approximation for speech as shown in Fig. 1 is given by:

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

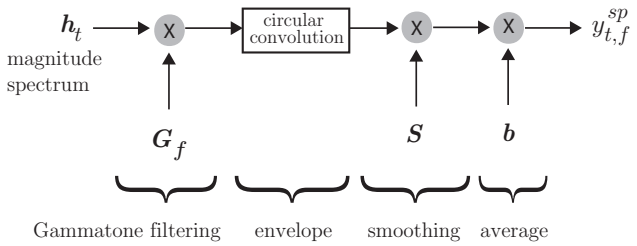

Fig. 1. Proposed approximation for the Spectro Temporal Excitation Pattern (STEP) calculation for speech.

where N is the number of frequency bins of the spectrum, $\circledN$ is the circular convolution operation dimension N and:

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

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

$\boldsymbol{S} = \text{diag}\left(\left[\, \upsilon_1 \;\; \ldots \;\; \upsilon_N \,\right]\right)$ is an NxN diagonal matrix whose diagonal contains the frequency response of the smoothing filter;

$\boldsymbol{b} = \left[\, b_1 \;\; \ldots \;\; b_N \,\right]$ is an Nx1 vector containing the coefficients of average filter.

The approximated version of the GP measure proposed here obtains correlation coefficients that are smaller but still comparable to the ones obtained by the original GP measure and higher than the ones obtained by any other spectrum-based measure when using the subjective data from the experiment described in [4].

4.2. Cost function reformulation

In order to keep a compromise between the minimization of the cost function defined in Eq. (2) and the maximization of the intelligibility measure given by Eq. (3) we need to define a parameter β that controls the weight given to each criterion. The redefined cost function is:

$$E_t = \varepsilon_t - \beta\, GP_t \tag{5}$$

where ε_t is the value of the function described in Eq. (2) in time frame t and GP_t is the time evolution of the GP as defined in Eq. (3):

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

The cepstral coefficient vector $\boldsymbol{c}_t = [c_{t,1} \;\; \ldots \;\; c_{t,m} \;\; \ldots \;\; c_{t,M}]^\top$ is given by:

$$\boldsymbol{c}_t = \text{argmin}\left[\varepsilon_t - \beta\, GP_t\right] \tag{7}$$

It is clear that when $\beta = 0$ the proposed cepstral extraction method becomes the original method of Section 2.

4.3. Solving the optimization problem using Steepest Descent

The update equation for cepstral coefficients using Steepest Descent is:

$$\boldsymbol{c}_t^{(i+1)} = \boldsymbol{c}_t^{(i)} - \mu \boldsymbol{\nabla} E_t^{(i)} \tag{8}$$

where μ is the step size and the i index refers to iterations. From now on we are dropping the i index for clarity reasons.

According to the definition of the error given by Eq. (5) the gradient vector is:

$$\boldsymbol{\nabla} E_t = \boldsymbol{\nabla} \varepsilon_t - \beta \boldsymbol{\nabla} GP_t \tag{9}$$

The formula expressing the value of $\boldsymbol{\nabla}\varepsilon_t$ can be found in [6]. Considering the definition of the STEP function and the GP_t as given by Eqs. (4) and (6) we have that:

$$\boldsymbol{\nabla} GP_t = \frac{100}{N_f N} \sum_{f=1}^{N_f} \eta\, \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \big[1 - \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})\big] \cdot \boldsymbol{H}_{c_t} \boldsymbol{G}_f \left(2\, \boldsymbol{\Gamma}_N \circledS \boldsymbol{G}_f \boldsymbol{h}_t\right) \boldsymbol{S}\, \boldsymbol{b} \tag{10}$$

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 $(\boldsymbol{\Gamma}_N \circledS \boldsymbol{G}_f \boldsymbol{h}_t)$ defines an NxN matrix of the following form:

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

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

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

$$\frac{\partial |H_t(\omega_j)|}{\partial c_{t,m}} = |H_t(\omega_j)| \cos(m\, \omega_j) \tag{11}$$

4.4. Energy normalization

In order to avoid the trivial solution of maximizing the number of glimpses by increasing the overall energy level and to see how much we can improve intelligibility given a fixed Signal to Noise Ratio (SNR) we need to make sure that the optimization does not change the total energy of the signal at each time frame.

We assume that the excitation signal has power one, with magnitude response constant over all frequency range for both voiced (single pulse) and unvoiced (white noise) segments. Under this assumption and considering that the cepstral extraction method does not modify the excitation signal we can say with the help of the Parseval theorem that in order to keep the energy in the time domain constant it is sufficient to keep the following constant:

$$\psi = \sum_{j=1}^{N} |H(\omega_j)|^2 \tag{12}$$

An alternative solution to explicitly adding a constraint to the optimization problem is to normalize the spectrum at each iteration so that the signal in that frame has fixed energy. For this solution the only term that needs changing in the gradient vector $\boldsymbol{\nabla} E_t$ is the one given by Eq. (11), that for $m \neq 0$ becomes:

$$\frac{\partial |H'_t(\omega_j)|}{\partial c_{t,m}} = |H'_t(\omega_j)| \Big(\cos(m\, \omega_j) - \frac{1}{\psi} \sum_{l=1}^{N} |H_t(\omega_l)|^2 \cos(m\, \omega_l) \Big) \tag{13}$$

where $|H'_t(\omega_j)|$ is the energy normalized magnitude spectrum. It is possible to prove that there is no need to update the first cepstral coefficient c_0 in this solution as the normalization operation updates c_0 at each iteration to a certain value regardless of an additional Δc_0 term.

5. EVALUATION

We conducted experiments with vocoded and synthetic speech. The results for synthetic speech can then show us the impact of the acoustic modelling on the effectiveness of the method.

5.1. Experimental conditions

The speech material we used to generate vocoded speech was the semantically unpredictable sentences (SUS) set from the Blizzard Challenge 2010. The samples were of a British male speaker sampled at 20kHz. To train the models we used 1000 other sentences from the same speaker also at 20kHz. The same sentences used to generate vocoded speech were used as test sentences for the HMM-generated synthetic speech. To generate vocoded and synthetic speech we used as synthesis filter the log spectrum approximation filter [6] with simple excitation as input.

We used the same set of spectral and excitation parameters to analyse natural speech for both the generation of vocoded speech and the training of the acoustic model. Using the proposed method we extracted 52 cepstral coefficients for different β values, including the $\beta = 0$ case for comparison. The periodogram was set to be the smoothed spectrum extracted using STRAIGHT [10]. We initialize the algorithm with the first $M+1$ values of the minimum phase cepstrum. The step size was set to $\mu^{(i)} = 1/||\boldsymbol{\nabla} E_t^{(i)}||$. We used both error convergence and maximum distortion as stopping criterion.

The acoustic model we used for synthetic speech was a hidden semi Markov model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values. We used one stream for the spectrum and three streams for the logF0. We used the Global Variance method [11] to compensate for the oversmoothing effect of the acoustical modeling.

For these experiments we added vocoded and HMM-generated synthetic speech to two different types of stationary noise, speech shaped noise (ssn) and high frequency noise (hf). Each noise type was added at a different SNR: 0 dB for ssn and and -20 dB for hf.

For the listening test we played all signals over headphones to participants in soundproof booths. Each individual sentence could be played only once before the participant had to type in what he or she heard. A total of eight native English speakers participated in the experiment with vocoded speech and other eight participants were assigned to the experiment with synthetic speech. Each participant heard twelve different sentences per noise type.

5.2. Results and discussion

Fig. 2 shows the Long Term Average Spectrum (LTAS) of vocoded speech generated using the original and the proposed method when noise is speech shaped and SNR is 0 dB. In the figure we can also see the LTAS of the noise. We can see that on average the proposed method reallocates energy mostly to the frequency range between 800 Hz and 4.8 kHz, the band where the auditory human system is more sensitive. The attenuation occurs mostly in the lower frequency regions below 800 Hz. For the high frequency noise the energy boost occurs in a similar region and we also observed some attenuation in the high frequency region, as this region is highly masked by noise.

We observed that the proposed method improves not only the approximated GP measure introduced above but the original GP measure as well. This improvement was observed for all noise types and for both vocoded and synthetic speech.

Fig. 3 shows the word accuracy rates obtained in the listening test with vocoded (left) and synthetic speech (right). Each group

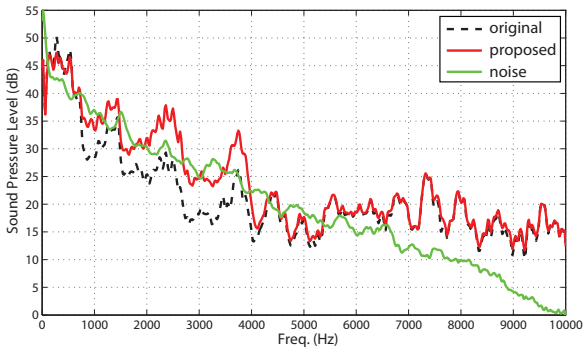

Fig. 2. Long term average spectrum curves extracted of vocoded speech generated using the original method ($\beta = 0$) and the proposed method ($\beta \neq 0$) for speech shaped noise at 0 dB SNR.

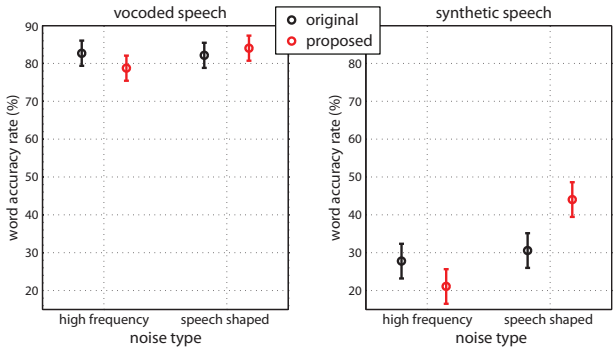

Fig. 3. Word accuracy rates of listening test with vocoded (left) and synthetic (right) speech.

mean is represented by a circle; two means are significantly different at a 0.05 level only if their intervals are disjoint.

We can see that the proposed method does not produce any significant differences in word accuracy for vocoded speech. However for synthetic speech and speech shaped noise there is a significant improvement of word accuracy from 31 % to 44 % (a gain of 44 %).

For the high frequency noise case it seems that, although not significantly, the proposed method decreases the word accuracy rates. We believe this happens because the modifications imposed by such noise leads to less natural speech which in turn degrades intelligibility. This could be solved by changing the acceptable amount of distortion and GP improvement or by stating the amount of distortion as a constraint instead of a stopping criterion.

The impact of the proposed method seems to be stronger for synthetic speech although the GP gains were smaller or similar for synthetic speech, most probably because in harder tasks smaller glimpse variations lead to stronger effects.

6. CONCLUSION

In this paper we showed how to use a measure of speech intelligibility in noise to modify HMM-synthetic speech and make it more intelligible for a certain noise. We proposed a new cepstral extraction method that aims not only to minimize the mismatch between periodogram and modelled spectrum but also to maximize speech intelligibility in noise, as defined by the Glimpse Proportion measure, given that the noise is known and SNR is known and fixed. The listening tests with vocoded and synthetic speech showed the effectiveness of the method for speech shaped noise but not for high frequency noise, which might indicate that the amount of distortion introduced into the speech by the modification was too large. Our next step is to handle distortion in a better way and then consider other types of constraints as well, for instance loudness. It is also in our plans to compare our approach to natural Lombard speech, in particular for those situations where humans are not fully able to change their own voice to successfully avoid the background noise.

Acknowledgment

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements 213850 and 256230 (SCALE and LISTA).

7. REFERENCES

[1] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, Sept. 2008, vol. 5.

[2] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, Kyoto, Japan, Sept. 2010.

[3] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Comm.*, vol. 51, no. 12, pp. 1253–1262, 2009.

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

[6] K. Tokuda, T. Kobayashi, and S. Imai, "Adaptive cepstral analysis of speech," *IEEE Trans. Speech and Audio Processing*, vol. SA-3, no. 6, pp. 481–489, Nov. 1995.

[7] S Imai and C. Furuichi, "Unbiased estimator of log spectrum and its application to speech signal processing," in *Proc. EURASIP*, Grenoble, France, Sep. 1988, pp. 203–206.

[8] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise," in *Proc. ICASSP*, Prague, Czech Republic, May 2011.

[9] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acta Acustica*, vol. 82, pp. 335–345, 1996.

[10] 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.

[11] 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.