Evaluating speech intelligibility enhancement for HMM-based 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

It is possible to increase the intelligibility of speech in noise by enhancing the clean speech signal. In this paper we demonstrate the effects of modifying the spectral envelope of synthetic speech according to the environmental noise. To achieve this, we modify Mel cepstral coefficients according to an intelligibility measure that accounts for glimpses of speech in noise: the Glimpse Proportion measure. We evaluate this method against a baseline synthetic voice trained only with normal speech and a topline voice trained with Lombard speech, as well as natural speech. The intelligibility of these voices was measured when mixed with speech-shaped noise and with a competing speaker at three different levels. The Lombard voices, both natural and synthetic, were more intelligible than the normal voices in all conditions. For speech-shaped noise, the proposed modified voice was as intelligible as the Lombard synthetic voice without requiring any recordings of Lombard speech, which are hard to obtain. However, in the case of competing talker noise, the Lombard synthetic voice was more intelligible than the proposed modified voice.

Index Terms: HMM-based speech synthesis, intelligibility of speech in noise, Lombard speech

1. Introduction

Many studies have shown that Lombard speech is more intelligible than speech produced in quiet when mixed with noise at the same SNR [1, 2, 3]. These studies and many others have also observed the following acoustic changes in Lombard speech: intensity increase, overall duration increase (vowels and consonants are affected in different ways however), increase in fundamental frequency, flattening of the spectral tilt (more energy is reallocated to higher frequencies). Although artificially narrowing F0 range has been found to decrease intelligibility of speech in noise [4], artificially increasing average F0 and F0 range [5] has been found not to increase intelligibility. It is still unclear to what extent the acoustic changes observed in natural Lombard speech improve intelligibility and how they are related to the characteristics of the noise.

In this work we are interested in increasing the intelligibility of Text-To-Speech (TTS) voices in a automated manner according to the environmental noise. For this, we make use of the versatile parametrical statistical HMM-based speech synthesis framework [6]. Under this framework we can use Lombard speech data, if available, to create a Lombard-like voice through the use of adaptation techniques [7] or in the case that such data are not available we can either modify the models or the generated sequence of acoustic parameters such that generated speech sounds Lombard-like, by increasing F0, flattening spectral tilt and increasing duration for instance [8]. However, to modify speech appropriately for the particular noise (type and SNR), we need to be able to automatically detect what sort of modifications should take place. To do that, we need a measure of intelligibility that can be used control which aspects of speech need to be modified and to what extent.

To date, the best objective measures for speech intelligibility in noise are not capable of providing such information exactly. These measures are based on the effective signal processing that takes place in the human auditory system and are therefore not appropriate for controlling modifications that will impact on other stages of auditory processing. In order to identify which aspects of speech need to be modified and whether objective measures of intelligibility are able to predict the impact of such modifications, we have performed a series of listening experiments with modified TTS samples in noise [9]. In these experiments, we found that modifications in the spectrum domain had significant positive impact on subjective intelligibility scores and that various objective measures were able to predict this effect to a greater or lesser degree. Since then, we have proposed a method for controlling the shape of the spectrum envelope by using an approximation of the Glimpse proportion measure [10] for the extraction of cepstral coefficients [11]. In a more recent method [12], we used the Glimpse measure as the optimization criterion for Mel cepstral coefficients modification. In this paper we now show the results of a more extensive listening experiment involving normal and Lombard natural speech, as well as synthetic speech created from normal speech, from Lombard speech, and

from normal speech followed by our proposed modification.

In Section 2 we present the method for Mel-cepstral modification based on the Glimpse proportion measure. In Section 3 we illustrate the acoustic properties of normal and Lombard speech, for both natural and synthetic samples, and for the modified synthetic voice, then give the results of the listening test.

2. Mel cepstral modification based on the Glimpse proportion measure

In order to increase intelligibility of HMM-based Text-To-Speech we proposed a method for modifying the Mel cepstral coefficients of generated synthetic speech [12] using the the Glimpse Proportion (GP) measure [10]. We use a previously-proposed approximation of the GP measure [11] as an optimization criterion for the Mel cepstral coefficient transformation.

Given a set of Mel cepstral coefficients and a noise signal, we want to obtain a new set of Mel cepstral coefficients $\boldsymbol{c}_t$ that maximizes GP_t (the approximated value of the GP) at time frame t :

$$\boldsymbol{c}_t = \operatorname{argmax} GP_t \tag{1}$$

$$GP_t = \frac{100}{N_f} \sum_{f=1}^{N_f} \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \tag{2}$$

where $\mathcal{L}(.)$ is a logistic sigmoid function that accounts for the identification of glimpses, $y_{t,f}^{sp}$ and $y_{t,f}^{ns}$ are the approximations of time frequency auditory representation of speech and noise as used in the GP measure and N_f is the number of frequency channels used to create this representation.

The auditory representation of speech is calculated from the spectrum envelope (in the form of Mel cepstral coefficients) by means of a Gammatone filterbank, envelope extraction and averaging over time. To obtain the auditory representation of the noise signal we perform this calculation over its short term Fourier transform. The modification is performed in a frame by frame basis and there is no reallocation of energy across time frames, only within frequency bins. To modify the Mel cepstral coefficients of a certain time frame, the optimization criterion uses the spectral content of noise and speech at the current time only and could therefore be applied in an online fashion if desired.

As the GP measure does not account for distortions that might occur when speech is modified, we restricted the amount of generated distortion by decreasing the frequency resolution of the modification. In other words, we modify only the first few Mel cepstral coefficients, representing the gross spectral envelope. We observed that better results are obtained when we modify only the first two coefficients [12], excluding the zero-th coefficient that accounts for the log-energy.

3. Evaluation

We evaluate here the performance of two natural voices – normal and Lombard speech recordings – plus three synthetic voices – normal, Lombard and modified. First we describe how we built the synthetic voices from the natural normal and Lombard speech data; then we provide a comparison of the acoustic properties observed in the five voices. After that, we give the results of a listening experiment with speech-shaped noise and a competing talker.

3.1. Speech material

To build the synthetic voices 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 a speech-modulated noise based on another male speaker [13], played over headphones at a absolute value of 84 dBA.

We built three different voices for this evaluation: TTS, TTSGP and TTSLomb. The normal speech dataset available for this particular speaker was not phonetically balanced, so instead of building speaker-dependent voices, we created then by starting with a high quality average voice model which was first adapted to the 2803 sentences of the normal speech database, comprising three hours of material, resulting in the voice denoted TTS. The Lombard voice TTSLomb was based on voice TTS, further adapted using 780 recorded sentences from the Lombard speech dataset, comprising 53 minutes of recorded material. Again, the reason for using adaptation was the lack of phonetic balance in the speech dataset. All acoustic features of the Lombard speech dataset i.e, Mel cepstral coefficients, logF0, duration and band aperiodicity were adapted, resulting in the TTSLomb voice. Given that Lombard speech data are not always readily available it would be advantageous to be able to obtain improved intelligibility by modifying an existing synthetic voice (trained with normal, non-Lombard speech). For this purpose we created the voice TTSGP by applying the previously described Mel cepstral coefficient modification method to synthetic speech utterances generated from the TTS voice. Duration, F0 and excitation parameters remained unmodified.

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 [14]. 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. We used a hidden semi-Markov model as the acoustic model and the Global Variance method [15] to compensate for the over-smoothing effect inherent in the

	speech (secs.)	F0 (Hz)	F0 range (Hz)	spectral tilt (dB/octave)	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 1: Acoustic properties of the two natural voices: Normal and Lombard and the three synthetic voices: TTS, TTSGP and TTSLomb (explained in the text).

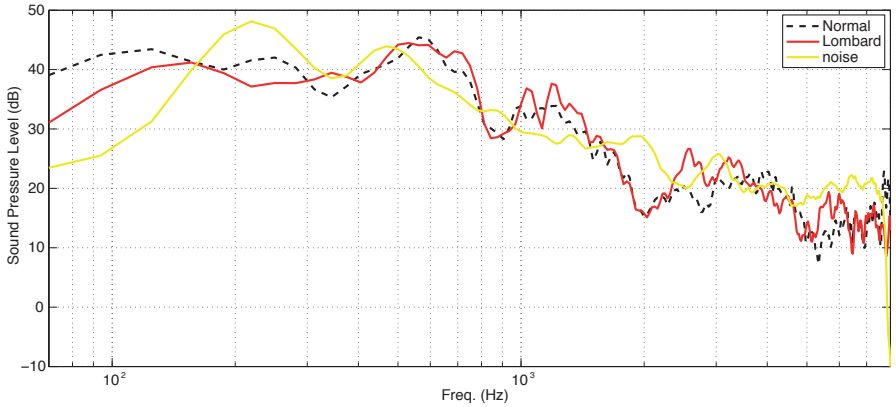

Figure 1: *Long term average spectrum of natural voices.*

acoustical modelling.

To modify the generated Mel cepstral coefficients, we used the method mentioned in the previous section. The auditory representation was constructed using 55 Gammatone filters that covered the range 50-7500 Hz as the noise signal used for testing was at a sampling rate of 16 kHz. Only the first two Mel cepstral coefficients were modified, excluding the log-energy coefficient; as the stopping criteria we used both error convergence and a distortion threshold set at a 10 % relative increase in the Euclidian distance between the auditory representation of original and modified speech.

3.2. Acoustic analysis

Table 1 shows various acoustic properties measured in the natural and synthetic data: duration changes of speech, prosody changes (in terms of average F0 values and range), spectral tilt and loudness, calculated using the ISO-532B method [16].

The natural Lombard sentences are on average 0.26 secs. longer than speech produced in quiet (a relative increase of 12%) and the synthetic Lombard TTSLomb sentences are 0.48 secs longer (which corresponds to a relative increase of almost 25%. The mean fundamental frequency value F0 is also higher for the Lombard voices, an increase of 27% and 39% for natural and synthetic speech respectively. The F0 range, calculated as the difference between the 80th and 20th percentile, also increases by 35% for natural and 90% for synthetic speech. Spectral tilt is found to be flatter: a relative change of 14% for natural and 24% for synthetic speech. The Lombard natural samples are on average 5% louder than normal speech ones and the Lombard synthetic voice TTSLomb is 13% louder than the normal synthetic voice TTS.

The voice built using the spectrum modification method TTSGP has the same duration and prosody as the TTS voice, but spectral tilt and loudness differ. The modified voice TTSGP presents a flatter spectral tilt when compared to the TTS voice (16% flatter) , though not to the same extent as the Lombard voice TTSLomb. The TTSGP is however slightly louder than the TTSLomb, a relative increase of 13% over the TTS voice.

The acoustic differences found here for the natural speech data are similiar to what has been reported in other studies of Lombard speech data: duration increases, F0 mean and range increases, flatter spectral tilt and increase in loudness. A similar but stronger trend was observed for the synthetic voices.

Figs.1 and 2 show the long term average spectrum

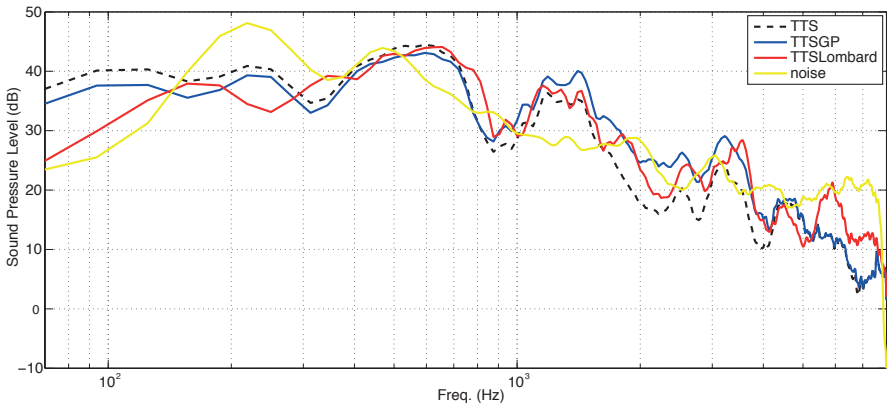

Figure 2: *Long term average spectrum of TTS voices.*

(LTAS) of a sentence from the natural speech recordings and from the generated synthetic speech respectively. In both figures we also display the LTAS of the noise that was used when creating the TTSGP voice: speech shaped noise presented at -4 dB. We can see from Fig.2 that, compared to the TTS curve, the curves for TTSLomb and TTSGP are attenuated at low frequencies, mostly below 1kHz and enhanced in the range above that. The TTSGP curve is mostly attenuated below 900 Hz and enhanced in the region between 900-4000 Hz. The TTSLomb voice presents similar behaviour but it seems less enhanced in this area and it is also enhanced in the region above 5 kHz. We can also see this effect in the natural Lombard speech curve displayed in the Fig.1, we can also see that the Lombard voices present a shift in fundamental frequency and formants.

3.3. Listening experiment

We mixed the five different speech datasets with two noise types: speech-shaped noise and speech from a single competing female talker. The noises were mixed at preselected signal to noise ratios (SNRs) chosen to achieve approximately 25, 50 and 75% word accuracy rates (-9 dB, -4 dB, 1 dB for speech shaped noise and -21 dB, -14 dB, -7 dB for competing talker).

The listening test involved 154 native English speakers listening to the noisy samples over headphones in sound-isolated booths. For the test, 180 sentences of the Harvard corpus were used in a balanced arrangement, such that listeners never heard the same sentence more than once. The subjective scores were computed from the word accuracy rates obtained per sentence and only content words were counted ('a', 'the', 'in', 'to', 'on', 'is', 'and', 'of', 'for' were excluded from scoring).

3.4. Subjective scores of intelligibility

Figs.3 and 4 show the word accuracy rates (WARs) of speech mixed with speech shaped noise (ssn) and competing talker (ct) for each SNR tested.

An obvious first comparison to draw is the difference in performance gain when using natural and synthetic Lombard speech. Averaged across the three different SNRs the gains in intelligibility obtained by the Lombard synthetic voice TTSLomb over the normal synthetic voice TTS are larger (47% for ssn and 42% for ct) than the gains obtained by the Lombard natural speech over the normal natural speech (17% for ssn and 13% for ct). That is, the Lombard effect was stronger in the synthetic voices. The effects are most pronounced for the lower SNRs cases for speech shaped noise and for the middle SNR case for the competing talker condition.

The noise played when recording the Lombard dataset used in this evaluation was different to the ones used in the listening test. We can thus infer that Lombard speech can still be more intelligible than speech produced in quiet even in a mismatched scenario. This would seem to indicate that certain modifications can provide improvements independent of the noise.

Most importantly we see that the post training modifications (TTSGP) also provide intelligibility gains over the non-Lombard synthetic voice (TTS). The word accuracy rates obtained by the TTSGP voice are comparable to those obtained with the TTSLomb voice for speech shaped noise even though no modification was made to duration or F0. Averaged across SNRs the relative gains obtained over the TTS voice were 44% for ssn and 5% for ct. For the competing talker only moderate improvements were obtained by TTSGP over TTS, suggesting a greater importance of prosody and duration in this scenario.

The TTS voices obtained lower WAR when compared to natural voices. On average across different noises and SNRs the TTS voice WAR is 23% lower than natural speech and TTSLomb WAR is 18% lower than the Lombard voice.

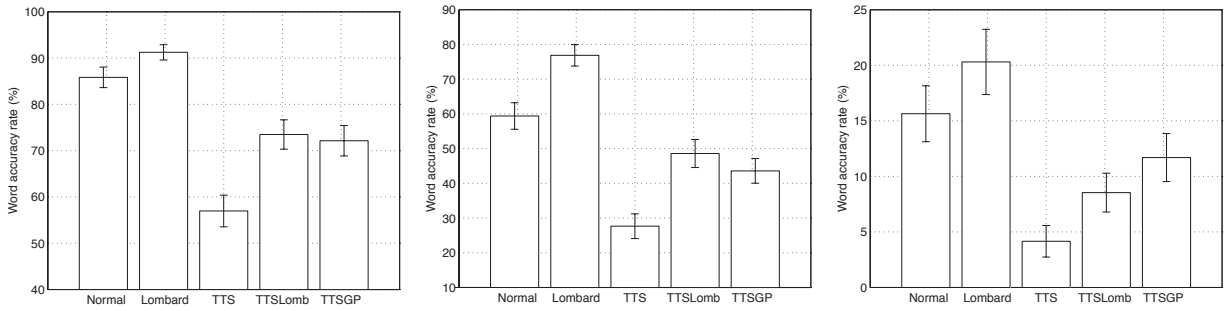

Figure 3: *Word accuracy rates for natural voices (Normal and Lombard) and synthetic voices (TTS, TTSLomb and TTSGP) mixed with speech shaped noise: SNR = 1dB (left), SNR =-4dB (middle) SNR =-9dB (right)*

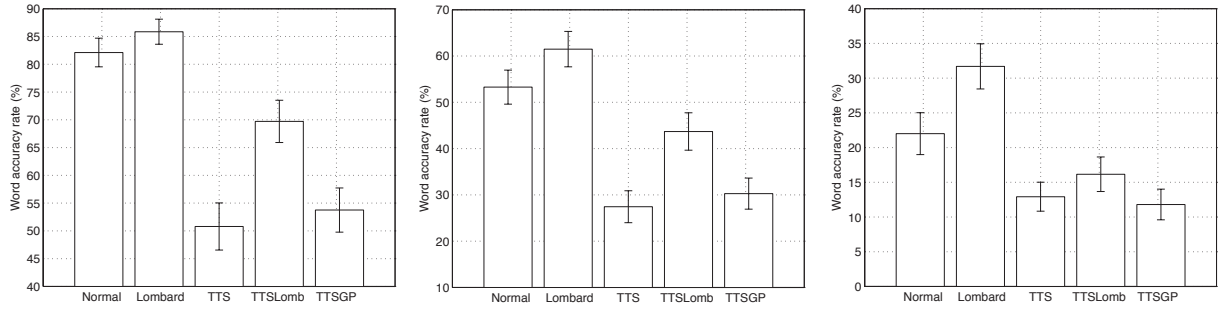

Figure 4: *Word accuracy rates for natural voices (Normal and Lombard) and synthetic voices (TTS, TTSLomb and TTSGP) mixed with a female competing talker: SNR = -7dB (left) SNR =-14dB (middle) SNR =-21dB (right)*

4. Conclusions

In this paper we have shown the results of an extensive listening experiment that evaluated the intelligibility of speech in noise. For this evaluation we compared synthetic voices built using speech produced in quiet and using Lombard speech (in a mismatched condition) as well as a modified synthetic voice whose spectrum envelope was automatically transformed depending on the noise signal. We found that the modified synthetic speech was as intelligible as the synthetic Lombard speech for a stationary noise (i.e., a purely energetic masker) but not in the presence of a competing talker. This can indicate that durational and prosodic changes are more important in the latter situation. We intend to extend our method for spectrum modification so that it is able to reallocate energy not only over frequency but over time.

5. 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.

6. References

[1] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[2] J. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, 1993.

[3] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3261–3275, 2008.

[4] J. S. Laures and K. Bunton, "Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions," *J. of Comm. Disord.*, vol. 36, no. 6, pp. 449 – 464, 2003.

[5] J. C. Krause, "Properties of naturally produced clear speech at normal rates and implications for intelligibility enhancement," PhD dissertation, MIT, Cambridge, MA, 2001.

[6] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[7] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66 –83, jan. 2009.

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

[9] 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.

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

[11] 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.

[12] C. Valentini-Botinhao, J. Yamagishi, and S. King, "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, September 2012.

[13] 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.

[14] 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.

[15] 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.

[16] ISO 532, "Acoustics - method for calculating loudness level," ISO, Geneva, Switzerland, 1975.