Analysis of HMM-Based Lombard Speech Synthesis

Tuomo Raitio[1], *Antti Suni*[2], *Martti Vainio*[2], *Paavo Alku*[1]

[1]Department Signal Processing and Acoustics, Aalto University, Helsinki, Finland
[2]Department of Speech Sciences, University of Helsinki, Helsinki, Finland
tuomo.raitio@tkk.fi, antti.suni@helsinki.fi

Abstract

Humans modify their voice in interfering noise in order to maintain the intelligibility of their speech – this is called the Lombard effect. This ability, however, has not been extensively modeled in speech synthesis. Here we compare several methods of synthesizing speech in noise using a physiologically based statistical speech synthesis system (GlottHMM). The results show that in a realistic street noise situation the synthetic Lombard speech is judged by listeners both as appropriate for the situation and as intelligible as natural Lombard speech. Of the different types of models, one using adaptation and extrapolation performed the best.

Index Terms: speech synthesis, HMM, Lombard effect, speech-in-noise

1. Introduction

When humans communicate in the presence of interfering noise, they tend to modify their voice in order to better deliver the message to the listener. This special voice is called the *Lombard speech* [1] or *speech in noise* [2]. In addition to increasing their loudness of speech, speakers also (involuntarily) modify the spectral qualities, durations, and prosody of speech in order to make the speech in noise more intelligible [1].

The intelligibility of speech and the Lombard effect have been studied extensively, but less in the context of speech synthesis. One reason for scarcity of previous work is that in concatenative synthesis paradigm, Lombard synthesis is hard to achieve as recording several hours of consistent Lombard speech is very difficult. Some studies have applied signal processing techniques to enhance the intelligibility of synthetic speech [2, 3], whereas in pure unit selection domain, one possibility is to measure an intelligibility score for units, to be used as an additional target cost factor, as was made in [4].

However, recent advances in hidden Markov model (HMM) based text-to-speech (TTS) have awakened increasing interests in speech-in-noise synthesis. HMM-based synthesis has provided good intelligibility results and has been applied successfully to various speaking styles. Furthermore, compared to, for example, emotional speech, Lombard speech has several attractive properties; Lombard effect is automatic and can be recorded without acting [2], it is text-neutral, and potential increase in intelligibility would benefit several real applications.

In Blizzard Challenge 2010 [5], we demonstrated that, in the presence of noise, synthetic speech can be made more intelligible than natural speech. However, the results were achieved by means of heavy signal modification, and the naturalness was compromised. Furthermore, the natural speech used as a reference in the challenge was recorded in a quiet studio and was not intended for noisy conditions. In this paper, we will investigate whether the high intelligibility results can be replicated while maintaining higher degree of naturalness by using Lombard speech as adaptation material. Secondly, we will compare synthetic Lombard speech to both conventional and Lombard style real speech. Thirdly, we will also consider the question of contextual appropriateness of speech in different noise conditions.

2. Modeling Lombard speech

There are several ways to model Lombard speech (e.g. unit selection/statistical synthesis, voice conversion). HMM-based text-to-speech (TTS) framework is especially attractive in the study of Lombard speech synthesis due to flexibility and limited training data requirements for modeling speaking styles through adaptation [6]. Thus, in this work, we have selected HMM-based synthesis as a platform. Consequently, three methods for creating synthetic Lombard voices were selected: (1) Modification of the synthesis vocoder, (2) Adaptation of the statistical models, and (3) Extrapolation of the adapted models. In the following section, the speech synthesizer and the Lombard modeling techniques are described.

2.1. Speech synthesis system

Our HMM-based speech synthesizers GlottHMM [7] is built on a basic framework of an HMM-based speech synthesis system [8], but it uses a special type of vocoder that attempts to model the speech production mechanism, with detailed parametrization of the voice source. In the parametrization stage, glottal inverse filtering [9] is utilized in order to separate the voice source and the vocal tract filter from the speech signal. The estimated vocal tract filter is described with line spectral frequencies (LSFs), while the voice source is described with the fundamental frequency (F_0), voice source spectrum (with LPC), and harmonic-to-noise ratio (HNR).

In synthesis stage, glottal flow pulses extracted from natural speech are used for creating the voiced excitation by interpolating the pulses according to F_0 and scaling in magnitude. The excitation is further modified according to the voice source parameters: The degree of voicing is controlled by matching the amount of noise in the excitation by manipulating the phase and magnitude of the spectrum of each pulse according to the HNR measure. Further, the spectral tilt of the excitation is modified by filtering the excitation by an adaptive IIR filter. White noise is used as unvoiced excitation. The vocal tract filter LSFs are enhanced [10] in order to alleviate for the over-smoothing, and the combined voiced and unvoiced excitation is filtered with the resulting vocal tract filter to create speech.

The GlottHMM system is described in detail in [7].

2.2. Creating synthetic Lombard speech by modification of the vocoder

In Blizzard Challenge 2010 [5], we created a special voice for the speech-in-noise tasks. In our synthetic voice, we modified phonetic features of speech and the vocoder of the synthesizer [11] according to several aspects of the Lombard effect. First, the rate of speech was lowered and the pitch was raised and its range was compressed. Maybe most importantly, the spectral tilt of voiced speech was decreased in order to model the speaker's effort to make the speech more audible. This concentrates more energy on the formant frequencies, where human auditory system is the most sensitive. The spectral tilt was decreased by modifying the spectrum of each glottal pulse. Finally, stronger post-filtering [10] was applied to produce a more prominent formant structure, and the speech waveform was companded in order the make the loudness of the speech signal as high as possible.

As a result, our speech-in-noise voice was the most intelligible of all the systems, even compared to natural speech [11]. However, the resulting speech quality was only average, but that was not evaluated in the speech-in-noise task. In this study, we have applied the same modifications as in Blizzard Challenge, but to address the speech quality degradation, the magnitude of the modifications was slightly toned down, using the Lombard adaptation corpus (see Sec. 2.3.1) as a reference.

2.3. Creating synthetic Lombard speech by adaptation

In HMM-based speech synthesis, trained speech models can be easily adapted with new training data. In this study, we wanted to find out how Lombard speech modeling in adaptation framework would compare to our method used in Blizzard Challenge and natural speech in terms of intelligibility and speech quality.

2.3.1. Recording speech-in-noise corpus

Three hundred short sentences, comprising approximately 30 minutes of speech by a Finnish male speaker, were recorded with AKG CK 92 omnidirectional condenser microphone with AKG SE 300 B preamplifier. Two hundred of the sentences were phonetically rich and the rest were chosen for prosodic interest. The recordings were performed in an anechoic chamber. The microphone was in 80 centimeters distance from mouth. Babble noise from NOISEX-92 database with 83 dB sound pressure level (SPL) was fed to the speaker's ears through Sennheiser HD 250 linear II headphones. The SPL of the noise through the headphones was measured with Cortex MK2 artificial head. Speaker's own voice was played back to headphones to control the degree of the Lombard effect. In the beginning, the loudness of the feedback was set to correspond to a situation in which the subject is speaking in a conventional quiet room.

2.3.2. Adapted models

In this study, we used CSMAPLR + MAP (constrained structural maximum *a posteriori* linear regression + maximum *a posteriori*) adaptation technique [6] to create a synthetic Lombard voice. Base voice was trained with 600 sentences spoken in normal style, and the 300 sentences of Lombard speech from the same speaker were used as the adaptation data. Adaptation was applied to all streams using state-tying decision trees for regression classes.

By informal listening, the Lombard effect produced by the adapted models was found somewhat weak. Thus, we decided to implement an additional "boosted" version of the adapted voice by means of interpolation. The possibility of interpolating between two or more model sets is a unique feature in HMM-based TTS, and it has been applied with success in producing a smooth continuum of speaking styles. Here, we took the advantage of the possibility of going beyond the voices to be interpolated, that is, to extrapolate. The extrapolation ratio between the normal voice and the adapted Lombard voice was set to 1.5 (where 0 would be normal voice and 1.0 the adapted voice). Extrapolation was applied to all streams, except for duration, where adapted models were used.

3. Evaluation

The aim of the evaluation is to assess various perceived characteristics of natural and synthetic speech in both conventional and Lombard speaking styles in different noise conditions. Naturally, the intelligibility of speech is of main interest, but the importance of quality is also addressed. Finally, we consider the question of contextual appropriateness of speech with respect to the noise environment.

3.1. Noise conditions

Since we are interested in real Lombard speech, a realistic noise environment was created. Real multichannel recording of street noise with most of the energy at low frequencies was chosen. First-order Ambisonics (B-format) recording was used, consisting of four channels: W, X, Y and Z. The W channel is the non-directional mono component of the signal captured with an omnidirectional microphone. The X, Y and Z channels are the directional components in three dimensions captured with three figure-of-eight microphones, facing forward, to the left, and upward. The 4-channel recording was rendered to the 9-channel loudspeaker setup by using directional audio coding [12].

Three noise levels were selected: silence, moderate noise, and extreme noise. Averaged A-weighted sound pressure levels for the noises were 63 dB for moderate and 70 dB for extreme noise. The noise levels were selected by pre-listening so that the differences in the intelligibility of different speech types were perceptually most prominent. The average SNRs were -1 dB and -8 dB for moderate and extreme noises, respectively.

3.2. Listening environment

All listening tests were performed in a standardized listening room that is in accordance with the recommendation ITU-R BS.1116-1. The listening room contained nine identical DSP-equalized loudspeakers (Genelec 8260A), with one speaker in front, two speaker on the front left and right, four speakers by the side of the listener in left and right, and two speakers behind the listener. The loudspeaker setup surrounded the listener creating a natural sounding and realistic sonic environment.

Noise was played to the listener through all the nine speaker, while speech was played only through the single front speaker, corresponding to a real speaking person.

3.3. Speech signals

In order to evaluate speech in noise, both natural and synthetic, the GlottHMM system was trained with a database of 600 sentences spoken by the same speaker as in the Lombard recordings (see Sec. 2.3.1). This baseline synthetic voice was compared with three types of synthetic Lombard voices: (1) Lombard speech synthesis by modification of vocoder [11], (2) Lombard speech synthesis by adapting the models with Lombard speech, and (3) Lombard speech synthesis by extrapolation of the Lombard adapted models. In addition, natural normal speaking style speech and Lombard speech were used in the test as references.

The recording of test sentences by the original speaker was performed in similar conditions as the speech-in-noise corpus (see Sec. 2.3.1). However, the speaker's own feedback was set very low, so the Lombard effect was significantly stronger than in the adaptation corpus. Thus, we acquired a challenging top line for the test.

The active speech level, or loudness, of all voice signals was normalized using the method in ITU-T P.56. After the normalization, we measured the averaged A-weighted sound pressure levels (SPL) of each voice type in the listening room. All the six voice types and their measured SPLs are shown in Table 1.

We were interested in a typical performance of the synthesizers, so instead of phonetically complex or semantically unpredictable sentences, ordinary short sentences were used. The sentence sets were designed to have a closely matching distribution of normal language, with regards to phoneme and lexical frequencies, word length, etc. The test sentences were manually annotated for word prominence in a four level scale.

A representative set of test voices is available online at http://www.helsinki.fi/speechsciences/synthesis/samples.html.

3.4. Listening tests

Two types of listening tests were performed in order to evaluate the differences between the voice types. First, in order to evaluate the intelligibility of speech, a specific intelligibility test was performed. In this test, short Finnish sentences were presented to the listener. The listener was allowed to listen to the samples only once, and was then asked to type in what they heard. Word error rates of the answers were evaluated, taking separately into account the inflectional and derivational suffixes. Second, the listener was asked to rate the samples according to four questions with specific verbal scales, both presented in Table 2. The listener was allowed to listen to the samples as many times as desired.

The sound samples were rated with continuous scales ranging from 0 to 100. Every listener rated five sentences of each six speech type in every three noise conditions. Thus each listener rated a total of 90 test samples in both test types (intelligibility and subjective rating). A total of 17 Finnish listeners (16 male and 1 female) with normal hearing took part in the test.

3.5. Results

The word error rates (WER) of the intelligibility test with 95% confidence intervals are shown in Fig. 1. In silence, all voice types are statistically equally intelligible with WERs ranging from 0.3% to 1.2%. In moderate noise, the WERs of all Lombard voices (*adapt* 4.4%, *vocod* 4.4%, *extrp* 4.8%, *lombd* 3.8%) are statistically equal, whereas both conventional speaking style voices (*convt* 13.3%, *speak* 17.1%) have much lower intelligibility. In the case of extreme noise, two synthetic Lombard voices (*vocod*, 39.1%; *extrp* 40.0%) are statistically as intelligible as natural Lombard speech (*lombd* 34.1%), while the third Lombard voice (*adapt* 52.2%) is less intelligible. Conventional speech (*speak* 91.1%) and synthesis (*convt* 93.2%) are almost totally unintelligible.

Table 1: *Test voice types and their averaged A-weighted SPLs after loudness normalization with ITU-T P.56.*

Type	Description	SPL
adapt	Lomb. synthesis by adaptation	62 dB
vocod	Lomb. synthesis by mod. of the vocoder	63 dB
extrp	Lomb. synthesis by extrapolation	63 dB
convt	Conventional speaking style synthesis	61 dB
speak	Natural conv. speaking style speech	59 dB
lombd	Natural Lombard speech	63 dB

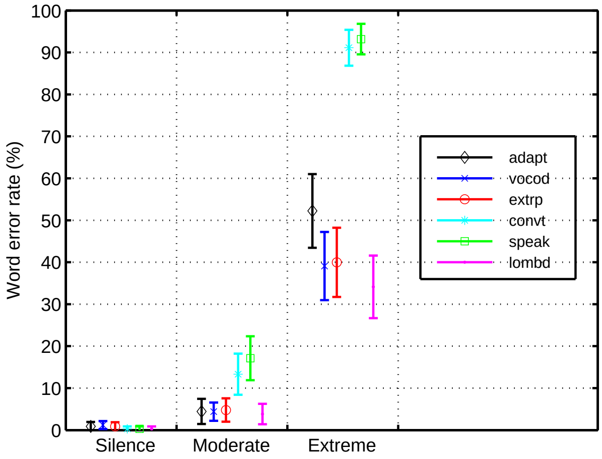

Figure 1: *Results of the intelligibility test in three noise conditions: silence, moderate (63 dB, SNR $= -1$ dB), and extreme (70 dB, SNR $= -8$ dB) street noise.*

The results of the subjective rating for the three noise conditions are shown in Fig. 2, 3, and 4 with 95% confidence intervals. In silence, both natural voices (*speak*, *lombd*) are rated to be of much higher quality compared to synthetic ones, but this gap vanishes in noisy conditions. The vocoder modified voice (*vocod*) is consistently rated worse in quality than the adapted voices. The suitability scores of both conventional speaking style voices (*convt*, *speak*) in silence are high (over 90), whereas all Lombard styles are rated less suitable, with natural Lombard speech (almost shouting) rated least suitable (around 50).

In noise conditions, the suitability scores are completely opposite, Lombard voices being more appropriate than conventional voices. In both noise conditions, all the Lombard voices are significantly different from conventional speaking style voices in terms of intelligibility, required effort and suitability. Subjective ratings of effort and intelligibility are very well in accordance with the true intelligibility scores.

4. Conclusions

The results of the evaluation show that the synthesized Lombard voices are generally more intelligible and more suitable to be heard in the presence of noise compared to conventional voices.

Table 2: *Questions and verbal scales of the subj. evaluation.*

How much effort did it take to understand the sentence? *very little – little – some – much – very much*
How well did you understand the sentence? *badly – poorly – fairly – well – very well*
How would you rate the quality of the speech sample? *bad – poor – fair – good – excellent*
How suitable was the speaking style considering the sonic environment? *badly – poorly – fairly – well – very well*

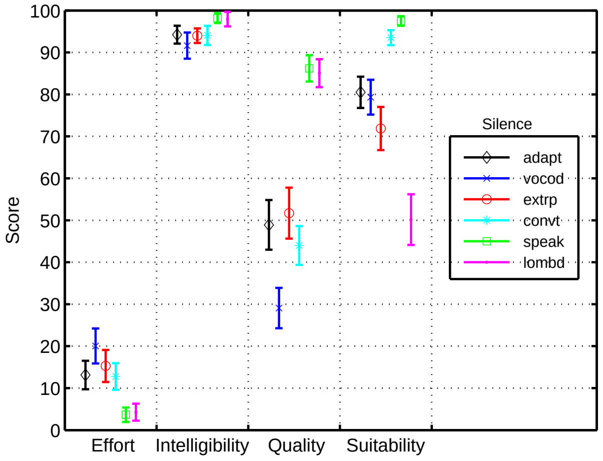

Figure 2: *Results of the subjective evaluation in silence.*

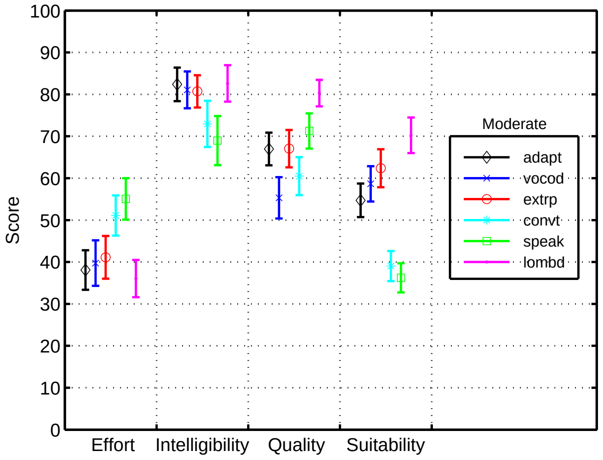

Figure 3: *Results of the subjective evaluation in the presence of moderate street noise (average SNR* $= -1$ *dB).*

They are rated very similarly as natural Lombard speech. The adapted Lombard voices are considered of higher quality than the Lombard voice generated by vocoder modification.

The used loudness normalization method ITU-T P.56 (also used in Blizzard Challenge) produced slightly different average A-weighted SPLs for different voices. Naturally, Lombard speech has more energy on the higher frequencies compared to normal speech, and thus their A-weighted SPLs are higher. The question remains, as to what kind of loudness normalization method would be suitable for the evaluation of intelligibility of different voices and speaking styles. Furthermore, only male speech and street noise was used in this study. The low-pass noise masked the low-pitched male speech fairly effectively, whereas different noise conditions could have lead to somewhat different results. The main conclusions, however, would probably be the same.

In conclusion, outperforming the intelligibility of natural speech in noisy conditions is achievable with HMM-based synthesis, both by vocoder manipulation and adaptation. Adaptation combined with extrapolation performed best overall, even in terms of quality. The vocoder manipulation technique was not much worse in high noise conditions, and it can be useful if no adaptation data is available. Also, the degradation of speech quality caused by statistical modeling and vocoding loses its significance in the presence of noise.

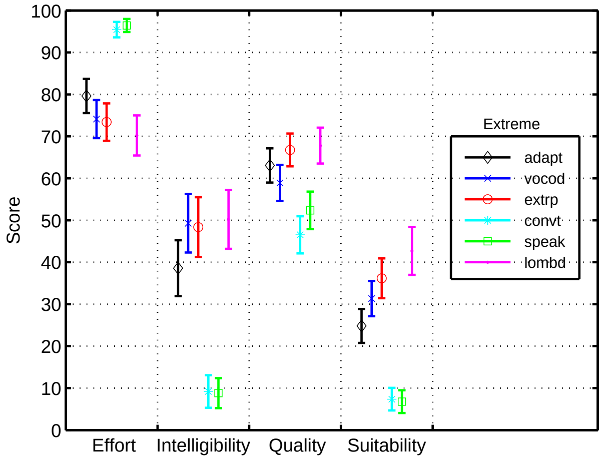

Figure 4: *Results of the subjective evaluation in the presence of extreme street noise (average SNR* $= -8$ *dB).*

5. Acknowledgements

This research is supported by the Academy of Finland (135003, 1128204, 1218259, LASTU), MIDE UI-ART, and Tekes. The authors would also like to thank E. Jokinen, M.Sc. M.-V. Laitinen, and M.Sc. H. Pulakka of Aalto University for their help.

6. References

[1] Van Summers, W., Pisoni, D., Bernacki, R., Pedlow, R. and Stokes, M., "Effects of noise on speech production: Acoustic and perceptual analyses", J. Acoust. Soc. Am., 84(3):917–928, 1988.

[2] Langner, B. and Black, A.W., "Improving the understandability of speech synthesis by modeling speech in noise", ICASSP 2005, pp. 265–268, Philadelphia, USA, 2005.

[3] Cernak, M., "Unit selection speech synthesis in noise", ICASSP 2006, pp. 14–19, May 2006.

[4] Patel, R., Everett, M. and Sadikov, E., "Loudmouth: Modifying text-to-speech synthesis in noise", 8th intl. ACM SIGACCESS conf. on Computers and Accessibility, New York, NY, 2006.

[5] King, S. and Karaiskos, V., "The Blizzard Challenge 2010", The Blizzard Challenge 2010 workshop, 2010. Online: http://festvox.org/blizzard

[6] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. and Isogai, J., "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm", IEEE Trans. on Audio, Speech, and Lang. Proc., 17(1):66–83, Jan. 2009.

[7] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE Trans. on Audio, Speech, and Lang. Proc., 19(1):153–165, Jan. 2011.

[8] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W. and Tokuda, K., "The HMM-based speech synthesis system (HTS) version 2.0", Sixth ISCA Workshop on Speech Synthesis, 294–299, Aug. 2007.

[9] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech Commun., 11(2–3):109–118, 1992.

[10] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., "Comparison of formant enhancement methods for HMM-based speech synthesis", Seventh ISCA Workshop on Speech Synthesis, pp. 334–339, Kyoto, Japan, Sep. 2010.

[11] Suni, A., Raitio, T., Vainio, M. and Alku, P., "The GlottHMM speech synthesis entry for Blizzard Challenge 2010", The Blizzard Challenge 2010 workshop, 2010. Online: http://festvox.org/blizzard

[12] Pulkki, V., "Spatial sound reproduction with directional audio coding", J. Audio Eng. Soc. 55(6):503–516, Jun. 2007.