Analysis and synthesis of shouted speech

Analysis and Synthesis of Shouted Speech

Tuomo Raitio[1], *Antti Suni*[2], *Jouni Pohjalainen*[1], *Manu Airaksinen*[1], *Martti Vainio*[2], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[2]Department of Behavioural Sciences, University of Helsinki, Helsinki, Finland

Abstract

In this study, the acoustic properties of shouted speech are analyzed in relation to normal speech, and various synthesis techniques for shouting are investigated. The analysis shows large differences between the two styles, which induces difficulties in synthesis. Analysis-synthesis experiments show that the use of spectral estimation methods that are not biased by the sparse harmonics of shouted speech is beneficial. The synthesis of shouting is performed through adaptation and voice conversion. Subjective evaluation of synthesis reveals that, despite quality degradation, the impression of shouting and use of vocal effort is fairly well preserved. In addition, the use of specific spectral estimation methods is found to be beneficial also in adaptation.

Index Terms: shouting, speech analysis, speech synthesis

1. Introduction

Shouting is the loudest mode of vocal communications, which is usually used for increasing the signal-to-noise ratio (SNR) when communicating over a distance or over an interfering noise. Alternatively, shouting can be used for expressing emotions or intentions. Shouting is different from Lombard speech [1] or speech with increased vocal effort; it lies at the extreme end of vocal effort continuum and is a voluntary phenomenon. It is a compromise between increased sound pressure level (SPL) and intelligibility; while loudness-normalized Lombard speech or speech with increased vocal effort is usually more intelligible than normal speech in the presence of noise [2, 3], extreme shouting is less intelligible [2, 3].

In addition to a large increase in SPL, shouting is characterized by a higher fundamental frequency (F0) due to increased subglottal pressure and vocal fold tension [4]. The F0 contour also shows less variety as F0 tends to saturate towards high values due to physical constraints, due to which F0 contours of different speakers also become similar to each other [4]. The relative length of the glottal closing phase decreases as the vocal effort is increased [5]. In the frequency domain, this increased sharpening of the glottal pulse in the time domain results in the emphasis of higher frequencies. In shouting, voiced sounds tend to elongate, unvoiced sounds become shorter, and the vowel-to-consonant ratio increases. Shouted speech is also less articulated; especially the first formant, and also the second, are shifted so that vowels become more similar to each other [6]. Due to the less accurate articulation and relative decrease in consonant energy, shouted speech is less intelligible than normal speech [2].

The properties of shouted and normal speech are studied e.g. in [2, 4, 6, 7], and the intelligibility of shouted and normal speech in the presence of noise is investigated e.g. in [2, 3]. Classification and detection of shouted speech is studied e.g. in [8, 9]. The effect of increased vocal effort on speech has also been widely studied [10, 11, 12]. However, there are very few studies that explicitly address the *synthesis* of shouted speech. Previous studies concentrate on increased vocal effort in concatenative speech synthesis [13, 14, 15, 16] or in voice conversion [17, 18]. In the context of statistical parametric speech synthesis [19], many studies concentrate on expressive speech synthesis, but vocal effort is explicitly modeled only in [20, 21, 22], and no studies can be found on synthesis of shouted speech.

The aim of this paper is to study various methods to create synthetic shouted speech. The study is based on first recording and analyzing shouted and normal speech, described in Sec. 2 and 3. Based on the analysis results, various synthesis techniques for creating shouted speech are experimented with. The methods and evaluation results are described in Sec. 4. Finally, Sec. 5 summarizes the current findings.

2. Data

Normal and shouted speech was recorded from 11 male and 11 female native speakers of Finnish. 24 sentences were recorded both in normal phonation and shouting. 12 of the sentences are in the imperative mood, consisting of one to four words. The semantic contents of these sentences represented vocal messages that people might use in potentially threatening situations. The other 12 sentences, each consisting of three words, are in the indicative mood and have a neutral, abstract information content. Recordings were performed in an anechoic room with AKG CK92 omnidirectional capsule at 70 cm distance from the speaker with SE300B power supply. Speech was sampled at 96 kHz with 24-bit resolution, after which the data was downsampled to 16 kHz. A calibration signal (1 kHz, 92.3 dB) was also recorded to determine the actual SPL of recorded speech.

First, speech of normal vocal effort was recorded, after which the speakers repeated the sentences using shouted voice. Speakers were instructed to use a very large vocal effort when shouting, which was controlled both by listening the recording and monitoring the signal waveform on the computer. If the intensity of shouting was not adequate, the talker was asked to repeat the sentence. A total of 528 sentences were recorded both in normal and shouted voice.

Shouted speech from two additional speakers, whose normal speech databases already existed, was also recorded. The normal databases consist of 599 and 1319 sentences spoken by a male and a female Finnish speaker. Both databases are designed for hidden Markov model (HMM) based speech synthesis purposes. Both speakers shouted 100 new sentences, of which 30 were three-word sentences with varying focus conditions, and the rest were short prose quotations with emotional content.

3. Analysis of normal and shouted speech

The recorded speech files were analyzed in terms of duration, SPL, F0, the difference between the first and the second har-

Table 1: Mean and 95% confidence intervals of speech features for female and male normal and shouted speech.

Feature	Unit	Female normal	Female shout	Male normal	Male shout
Duration	s	1.35 ± 0.03	1.62 ± 0.04	1.42 ± 0.04	1.76 ± 0.05
SPL	dB	61.6 ± 0.1	79.2 ± 0.1	63.0 ± 0.1	82.7 ± 0.1
F0	Hz	209.9 ± 0.3	359.7 ± 0.7	102.4 ± 0.2	259.4 ± 0.6
H1–H2	dB	11.56 ± 0.05	9.26 ± 0.07	9.01 ± 0.04	9.44 ± 0.06
NAQ	-	0.0729 ± 0.0003	0.0563 ± 0.0002	0.0607 ± 0.0002	0.0599 ± 0.0002

monic (H1–H2) [23], and the normalized amplitude quotient (NAQ) [24]. The reported SPLs were calibrated to correspond to the actual SPLs in the recordings. Duration was analyzed per sentence while the rest of the features were analyzed frame-wise using the GlottHMM vocoder [25]. The results of the analysis are shown in Table 1.

Analysis results show that the duration of shouted sentences were 20% and 24% longer compared to normal speech for female and male speakers, respectively. The SPL of shouting was on average 17.6 dB and 19.6 dB higher than that of normal speech for females and males, and the mean F0 for shouting was 71% and 152% higher compared to normal speech for females and males, respectively. H1–H2 was on average 2.3 dB lower in shouting than in normal speech for females, indicating decreased spectral tilt in shouting. However, for males H1–H2 was 0.4 dB higher in shouting than in normal speech. NAQ was 23% and 1% lower for females and males, respectively, indicating tenser phonation type in shouting for both genders.

The frame-wise distributions of SPL, F0, H1–H2, and NAQ for normal and shouted speech are shown in Fig. 1. The energy-normalized spectra of normal and shouted speech are shown in Fig. 2, which shows how the vowel-to-consonant ratio increases dramatically in shouted speech. The time and amplitude normalized average female and male glottal flow derivative waveforms of normal and shouted speech are shown in Fig. 3. For both genders, the pulse waveform in shouting shows a shorter open phase and a more abrupt glottal closure.

Despite the rigorous recording arrangements, the properties of shouting vary considerably between speakers. The average shouting SPLs vary over a 17 dB range both in the female and male speaker's groups. This is probably due to at least two reasons. First, speakers have different conception of shouting; some consider it as a loud voice, other consider it as an extreme vocal expression in which the vocal apparatus is on its limits. Second, it is difficult to elicit extreme shouting in laboratory conditions; speakers may find it embarrassing or inconvenient to use their extreme voice. Despite the personal variation in shouting, the SPL difference between shouted speech and normal speech ranged from 17 dB to 28 dB for the female speakers and from 15 dB to 33 dB for the male speakers. This is in line with previous studies: Rostolland [4] reports C-weighted level differences between shouted and normal speech of 20 dB and 28 dB for female and male speakers, respectively.

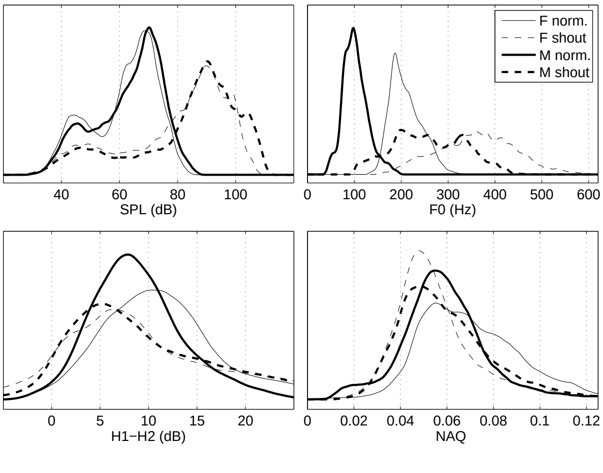

Figure 1: *Distributions of SPL, F0, H1–H2, and NAQ for female (F) and male (M) normal and shouted speech.*

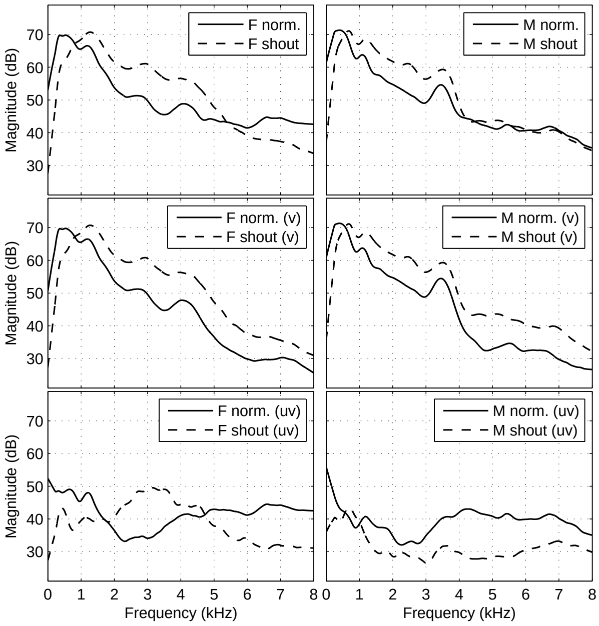

Figure 2: *Average spectra of energy-normalized normal speech and shouting for female (F) and male (M) speakers. Top graphs represent the overall spectra, middle graphs only voiced spectra (v), and bottom graphs only unvoiced spectra (uv).*

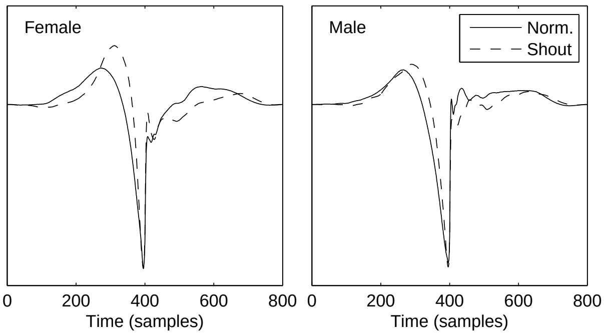

Figure 3: *Normalized average two-period glottal flow derivative waveforms of female and male normal and shouted speech.*

4. Synthesis of shouted speech

The ultimate aim of speech synthesis is to artificially create any type of vocal expression. Shouting is a rarely used but very distinct and important type of vocal expression. Although SPL of a sound can be increased simply by amplifying the signal waveform, synthesis of natural sounding shouted speech calls for adjusting a number of acoustical features as discussed in Sec. 1. Such modifications are required e.g. in creating speech with emotional content, which finds use in human-computer interaction, creating virtual agents, and communication aids. In the following, various techniques for synthesizing natural-sounding shouted speech are experimented with.

4.1. Analysis-synthesis

Compared to normal speech, shouting is characterized by a very high F0. This imposes difficulties in estimating the formant structure of speech because the sparse harmonic peaks may distort the estimation of the correct formant frequencies. Linear prediction (LP), commonly used for estimating speech spectrum, is especially sensitive to such errors. In order to avoid this problem, e.g. weighted linear prediction (WLP) [26] can be used to de-emphasize the effect of the excitation to formants. In this paper, WLP is experimented with two different weighting functions: short-time energy (STE) function [26] with stabilization (SWLP) [27] and attenuation of the main excitation (AME) [28]. In STE weighting, excitation is attenuated by a window defined by the short-time energy of the frame, while in the AME weighting the excitation is attenuated during the main excitation, defined by the glottal closure instants (GCIs).

Experiments are performed with the GlottHMM vocoder [25], which is a physiologically oriented vocoder that utilizes glottal inverse filtering [29] for speech analysis and natural glottal flow pulses for synthesis. It has been shown to successfully synthesize e.g. Lombard speech [22]. The quality of vocoder analysis-synthesis was experimented with the three spectral estimation methods: (1) conventional LP, (2) stabilized WLP with STE window (denoted as STE), and AME-WLP (denoted as AME). An ABX listening test comparing the analysis-synthesis quality of these methods was conducted with normal and shouted speech. In the test, listeners were presented with a natural reference sample and two vocoded samples. The task of the listener was to select the one that sounded more like the reference sample (or no preference). 15 native Finnish listeners each assessed a total of 60 sample pairs, consisting of 20 samples from each method. The results, shown in Fig. 4, indicate that AME-WLP performs best with normal speech, while stabilized STE-WLP performs best with shouted speech.

4.2. Conversion from normal to shouted voice

Voice conversion from normal speech to shouting is another way to artificially create shouted speech. Voice conversion is especially useful in the case where there is not enough shouted speech for training or adapting a statistical parametric speech synthesizer. In this paper, a simple voice conversion method is experimented with. The method is implemented in the pulse library version of the GlottHMM vocoder [30]. First, the vocoder is used to extract a database from the available speech data of the desired style (e.g. shouting), which comprises voice source and vocal tract parameters and extracted glottal flow pulses. In voice conversion stage, the speech parameters of normal voice, generated by a HMM-based synthesizer, are fed into the vocoder that adapts the means and variances of the source and filter parameter trajectories and uses the pulses in the database to construct the voice source signal. The parameters in the database are F0, energy, harmonic-to-noise ratio (HNR), voice source spectrum, and the vocal tract filter. In addition, utterance durations are uniformly stretched by 20% to match the lengthening in shouted speech. Although more sophisticated voice conversion techniques exist, this simple technique is used as a baseline in the evaluation (see Sec. 4.5).

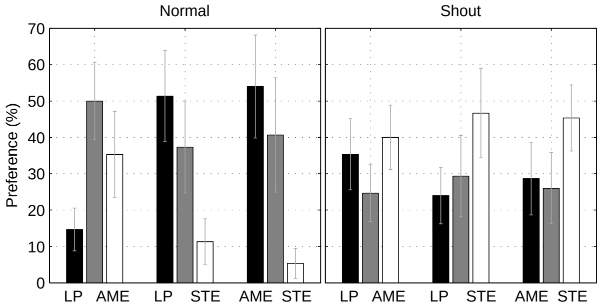

Figure 4: *Results of the ABX test comparing LP, AME-WLP, and stabilized STE-WLP for normal and shouted speech. The middle gray bar depicts no preference for either of the methods.*

4.3. HMM-based synthesis

In concatenative speech synthesis [31, 32], building a voice with a specific speech style requires a large database of style-specific speech for covering sufficient speech units. Especially with shouted speech, creating a large database with constant quality is impractical. Statistical parametric speech synthesis (or HMM-based speech synthesis) [19] provides an easy and flexible framework for synthesizing voices with different styles. A rather small database of normal speech can be used for training the base voice, which can be then adapted [33] to any voice type using only a small amount of speech with the specific style.

In this study, such an adaptation scheme is used to create synthetic shouted speech. Two normal speaking style voices, a female and a male (see Sec. 2), were built with the standard HTS method [34], accommodated to the extended stream structure of the GlottHMM vocoder [25], Then, the voices were adapted using the 100 sentences of shouted speech with CSMAPLR + MAP method [33], which was tuned for previous Lombard speech synthesis experiments in [22]. Two adapted shouting voices were built for both speakers, one with conventional LP and another with stabilized STE-WLP parameterization of the vocal tract spectrum. For normal speaking style voices, the vocal tract was parametrized with conventional LP.

Phonetically and lexically balanced set of 32 test sentences were generated with both normal and shout-adapted voices. Global variance [35] was not considered with the female voices due to observed artefacts in the shouted LP-based voice, but moderate amount of post-filtering [36] was applied to compensate for the over-smoothing of the formants.

4.4. Effect of spectral estimation method to adaptation

Analysis-synthesis experiments with shouted speech showed that conventional LP is prone to biasing the spectral estimate towards harmonics, thus creating artefacts. The same phenomenon is likely to happen when adapting the normal speech material with shouted speech. Thus, stabilized STE-WLP, which performed best for shouted speech, and LP are compared in terms of the quality of adapted shouted speech. A comparison

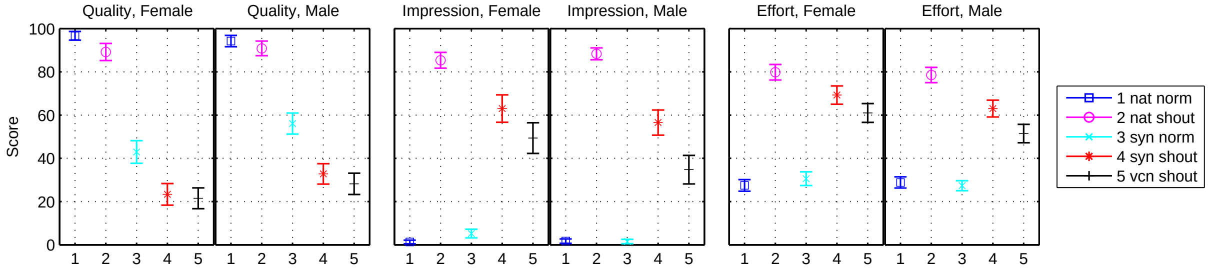

Figure 5: *Subjective evaluation results (quality, impression of shouting and effort) for natural and synthetic normal and shouted voices.*

category rating (CCR) test was conducted to find out if spectral estimation method has effect on the quality. In the CCR test, subjects are presented with speech sample pairs and the task of the listener is to rate the quality difference between the samples on the comparison mean opinion score (CMOS) scale, which is a seven-point scale ranging from much worse (−3) to much better (3). 40 sentences (5 for each speech type) were presented to each 11 listeners. The responses of the CCR test are summarized by calculating the mean score for each method with 95% confidence intervals, which yields the order of preference and distances between the methods. The results of the CCR test, shown in Fig. 6, indicate that the stabilized STE-WLP is preferred over LP especially with the high-pitched female voice.

4.5. Evaluation

A listening test was conducted in order to find out how synthetic shouted speech is perceived in comparison to natural normal speech, natural shouting, and synthetic normal speech. In addition, a voice conversion from HMM-based normal speech to shouted speech, described in Sec. 4.2, was used as a baseline. Thus, five different types of speech were included in the test:

1. Natural normal speech
2. Natural shouted speech
3. Synthetic normal speech (LP)
4. Synthetic shouted speech (stabilized STE-WLP)
5. Synthetic normal speech + voice conversion to shouting

A mean opinion score (MOS) type test was used for evaluation. Listeners were presented with a speech sample at a time and asked three questions assessing the quality of the speech sample, amount of perceived shouting, and impression of the amount of vocal effort used by the speaker. The answers were given with a continuous slider guided by five-point verbal scales. The questions and verbals scales are shown in Table 2. The loudness of the speech samples were normalized according to ITU-T P.56 [37] so that listeners perceived shouting not due to increased sound intensity, but due to the other acoustics features generated by the synthesis technique. 11 native Finnish listeners participated in the test conducted in quiet listening booths with headphones. Each listener rated 50 speech samples composed of 10 random samples from each category.

The evaluation results, guided by the five-point descriptions, were first converted to scale from 0 to 100, after which means and 95% confidence intervals were computed for each speech type. The results of the listening test are shown in Fig. 5. The results indicate that the adapted shouting voice is rated inferior in quality compared to normal synthetic speech. This is expected since adaptation of normal speech with small and very different data produces artefacts to prosody and speech quality. However, the impression of shouting is fairly well preserved as well as the impression of used vocal effort. The simple voice conversion technique, although rated similar in quality to adapted shouting voice due to more consistent prosody, is deteriorated in the impression of shouting and vocal effort.

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

How would you rate the quality of the speech sample? *bad – poor – fair – good – excellent*
How much does the sample resemble shouting? *none – little – moderately – much – very much*
How much effort did speaker use for producing speech? *very little – little – moderately – much – very much*

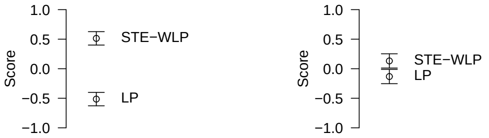

Figure 6: *Quality of shout adaptation with respect to the spectral estimation methods for female (left) and male (right) voice.*

5. Discussion and conclusions

Synthesis of shouted speech is challenging due to many reasons. First, it is hard to record a large database of shouting with consistent quality. Second, the difference between normal speech and shouting is very prominent. Shouting is characterized by high vocal energy and F0, increased duration, decreased spectral tilt, and reduced dynamics of formant frequencies. These changes induce problems in many speech processing algorithms. In this study, the biasing effect of high F0 to formant estimation was reduced by using specific spectral estimation methods.

HMM-based synthesis of shouting was experimented through adaptation and voice conversion. Subjective evaluation revealed that the quality of adapted shouting synthesis is degraded due to the aforementioned challenges: the large difference between the two styles and the small amount of adaptation data. These problems were most prominent in the prosody of synthesis. Stepwise adaptation from normal to shouted speech may improve the quality, which will be a topic of future work. However, the impression of shouting and use of high vocal effort is fairly well preserved. In contrast, voice conversion from synthetic normal speech to shouting exhibits more consistent prosody, but the characteristics of shouting are less prominent.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 287678 and from the Academy of Finland (256961, 135003).

7. References

[1] Lombard, E., “Le signe de l’elevation de la voix”, Ann. Maladies Oreille, Larynx, Nez, Pharynx, 37:101–119, 1911.

[2] Pickett, J., “Effects of Vocal Force on the Intelligibility of Speech Sounds”, J. Acoust. Soc. Am., 28(5):902–905, 1956.

[3] Rostolland, D., “Intelligibility of shouted voice”, Acustica 57(3):103–121, 1985.

[4] Rostolland, D., “Acoustic features of shouted voice”, Acustica 50(2):118–125, 1982.

[5] Alku, P., Airas, M., Björkner, E. and Sundberg, J., “An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity”, J. Acoust. Soc. Am., 120(2):1052–1062, 2006.

[6] Rostolland, D., “Phonetic structure of shouted voice”, Acustica 51(2):80–89, 1982.

[7] Elliott, J., “Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics”, Proc. SST-2000: 8th Int. Conf. Speech Sci. & Tech., 2000, pp. 154–159.

[8] Zhang, C. and Hansen, J., “Analysis and classification of speech mode: whispered through shouted”, Proc. Interspeech, 2007, pp. 2289–2292.

[9] Pohjalainen, J., Raitio, T., Yrttiaho, S. and Alku, P., “Detection of shouted speech in noise: human and machine”, J. Acoust. Soc. Am., 133(4), 2013 (accepted for publication).

[10] 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.

[11] Junqua, J.-C., “The Lombard reflex and its role on human listeners and automatic speech recognizers”, J. Acoust. Soc. Am., 93(1):510–524, 1993.

[12] Traunmüller, H. and Eriksson, A., “Acoustic effects of variation in vocal effort by men, women, and children”, J. Acoust. Soc. Am., 107(6):3438–3451, 2000.

[13] Schröder, M. and Grice, M., “Expressing vocal effort in concatenative synthesis”, Proc. 15th International Conference of Phonetic Sciences, 2003, pp. 2589–2592.

[14] Turk, O., Schröder, M., Bozkurt, B. and Arslan, L., “Voice quality interpolation for emotional text-to-speech synthesis”, Proc. Interspeech, 2003, pp. 797–800.

[15] Cernak, M., “Unit selection speech synthesis in noise”, Proc. ICASSP, 2006, pp. 14–19.

[16] Patel, R., Everett, M. and Sadikov, E., “Loudmouth: modifying text-to-speech synthesis in noise”, Proc. 8th Intl. ACM SIGACCESS Conf. on Computers and Accessibility, 2006.

[17] Langner, B. and Black, A.W., “Improving the understandability of speech synthesis by modeling speech in noise”, Proc. ICASSP, 2005, pp. 265–268.

[18] Huang, D.-Y., Rahardja, S. and Ong, E., ”Lombard effect mimicking”, Seventh ISCA Workshop on Speech Synthesis, 2010, pp. 258–263.

[19] Zen, H., Tokuda, K. and Black, A.W., “Statistical parametric speech synthesis”, Speech Commun., 51(11):1039–1064, 2009.

[20] 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

[21] Calzada, A. and Socoró , J., “Vocal effort modification through harmonics plus noise model representation”, In Advances in Nonlinear Speech Processing. Vol. 7015 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 96–103.

[22] Raitio, T., Suni, A., Vainio, M. and Alku, P., “Analysis of HMM-based Lombard speech synthesis”, Proc. Interspeech, 2011, pp. 2781–2784.

[23] Titze, I. and Sundberg, J., “Vocal intensity in speakers and singers”, J. Acoust. Soc. Am., 91(5):2936–2946, 1992.

[24] Alku, P., Bäckström, T. and Vilkman, E., “Normalized amplitude quotient for parametrization of the glottal flow”, J. Acoust. Soc. Am., 112(2):701–710, 2002.

[25] 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, 2011.

[26] Ma, C., Kamp, Y. and Willems, L., “Robust signal selection for linear prediction analysis of voiced speech”, Speech Commun., 12(1):69–81, 1993.

[27] Magi, C., Pohjalainen, J., Bäckström, T. and Alku, P., “Stabilised weighted linear prediction”, Speech Comm. 51(5):401–411, 2009.

[28] Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M. and Story, B., “Improved formant frequency estimation from high-pitched vowels by downgrading the contribution of the glottal source with weighted linear prediction”, Proc. Interspeech, 2012.

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

[30] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., “Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis” Proc. ICASSP, 2011, pp. 4564–4567.

[31] Hunt, A. and Black, A.W., “Unit selection in a concatenative speech synthesis system using a large speech database”, Proc. ICASSP, 1996, pp. 373–376.

[32] Black, A.W. and Campbell, N., “Optimising selection of units from speech database for concatenative synthesis”, Proc. Eurospeech, 1995, pp. 581–584.

[33] 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, 2009.

[34] 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, 2007, pp. 294–299.

[35] Toda, T. and Tokuda, K., “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis”, IEICE Trans. Inf. Syst. E90-D(5):816–824, 2007.

[36] 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, 2010, pp. 334–339.

[37] ITU, “Objective measurement of active speech level”, International Telecommunication Union, Recommendation ITU-T P.56, 2011.