Synthesis and Perception of Breathy, Normal, and Lombard Speech in the Presence of Noise

Tuomo Raitio[a,*], Antti Suni[b], Martti Vainio[b], Paavo Alku[a]

[a]*Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*
[b]*Department Behavioural Sciences, University of Helsinki, Helsinki, Finland*

Abstract

This papers studies the synthesis of speech over a wide vocal effort continuum and its perception in the presence of noise. Three types of speech are recorded and studied along the continuum: breathy, normal, and Lombard speech. Corresponding synthetic voices are created by training and adapting the statistical parametric speech synthesis system GlottHMM. Natural and synthetic speech along the continuum is assessed in listening tests that evaluate the intelligibility, quality, and suitability of speech in three different realistic multichannel noise conditions: silence, moderate street noise, and extreme street noise. The evaluation results show that the synthesized voices with varying vocal effort are rated similarly to their natural counterparts both in terms of intelligibility and suitability.

Keywords: Statistical parametric speech synthesis, Adaptation, Vocal effort, Lombard speech, Breathy speech, Intelligibility

1. Introduction

Humans adapt their vocal communication to the acoustic and auditory environment in order to successfully and efficiently deliver a message to a listener without using unnecessary effort. In environments with high levels of interfering noise, more effort is required in order to increase the signal-to-noise ratio (SNR) and thereby the intelligibility of speech. This automatic effect is known as the Lombard effect, and the speech produced in such conditions is called Lombard speech (Lombard, 1911) or speech-in-noise (Langner and Black, 2005). When speaking in silence or in low noise conditions, such an effort is not necessary, and the use of a softer voice is considered more appropriate. Thus, depending on the context, natural speech varies greatly from whispering or soft phonation to shouting. This variation in the use of vocal effort is called the vocal effort continuum.

Even though the vocal effort continuum is an integral part of human communication, it is typically not utilized in machine-to-human communication. In order to produce contextually appropriate synthetic speech, the auditory environment and context must be taken into account and speech must be produced at the corresponding point in the vocal effort continuum. The modeling of the vocal effort continuum in speech synthesis not only increases the intelligibility of the system in adverse conditions, but also makes the synthetic voice more natural and more appropriate for the listener (Raitio et al., 2011b). Thus, a message delivered through such a text-to-speech (TTS) system is more likely to be comprehended by the listener.

Conventionally, the modeling of the vocal effort continuum has been difficult in TTS due to the limitations of the techniques used. In concatenative speech synthesis (Hunt and Black, 1996; Black and Campbell, 1995), large amounts of speech data is required to cover sufficient units along the continuum. However, with statistical parametric speech synthesis techniques (Zen et al., 2009), the construction of such a continuum is relatively easy by using adaptation techniques (Yamagishi et al., 2009). One or a few smaller databases recorded along the continuum can be used to adapt the statistical parametric voice to any point on the continuum.

GlottHMM (Raitio et al., 2011c) is a statistical speech synthesis system that parametrizes speech by modeling the functioning of the real human speech production mechanism. Recently, GlottHMM was shown to enable synthesizing

*Corresponding author. Tel.: +358 50 4410733; fax: +358 9 460224. E-mail address: tuomo.raitio@aalto.fi.

rather natural sounding and highly intelligible normal and Lombard speech (Raitio et al., 2011b). In the current study, work on GlottHMM is extended by creating a continuum from low to high vocal effort using three databases along the continuum: breathy, normal, and Lombard speech. The intelligibility and contextual appropriateness are evaluated by synthesizing both female and male speech on the continuum in three realistic multichannel noise conditions: silence, moderate street noise, and extreme street noise. Compared to the authors' previous work on Lombard speech synthesis (e.g. Raitio et al., 2011b), the present investigation encompasses also the study of breathy speech, thus extending the vocal effort continuum to soft phonation. In addition, a more advanced synthesis technique is utilized, and the scope of the present study is also expanded by including female speech and a more extensive and versatile subjective evaluation.

The paper is organized as follows. The background of the study is discussed in Section 2 by addressing how vocal effort is reflected in different speech parameters and by describing previous investigations in the study area. Section 3 gives a detailed description of the method used in this study for creating synthetic voices along the vocal effort continuum. Subjective evaluation of the voices in realistic noise environment and the consequent findings are described in Section 4. Finally, Section 5 discusses the relevance and implications of the results and summarizes the paper.

2. Background

2.1. Properties of Vocal Effort

The acoustic properties of speech sounds change not only as a function of linguistic message, but also according to speaker, context, and expression. This study concentrates on the use and reproduction of different levels of vocal effort, which can depend on linguistic (Gordon and Ladefoged, 2001) and all of the three extralinguistic properties (Gobl and Ní Chasaide, 2003) mentioned above. Change in the vocal effort can be triggered by a noisy environment, in which case it is called the Lombard reflex or Lombard effect (Junqua, 1993). Change in the vocal effort can also be triggered by the need to communicate over a distance (Traunmüller and Eriksson, 2000) or a change in emotional expression (Ishi et al., 2010; Gobl and Ní Chasaide, 2003). This study will mainly address the case where the noise environment is the cause for the vocal effort changes, but also the effect of distance is discussed.

The effects of vocal effort on speech has been widely studied (see e.g. Rostolland, 1982; Summers et al., 1988; Junqua, 1993; Traunmüller and Eriksson, 2000). In high vocal effort, word duration is reported to be longer, the vowel duration increased, and the consonant duration generally decreased compared to normal speech. The mean fundamental frequency (f_0) of speech is also increased and its variance is decreased in high vocal effort. The formant frequencies are shifted due to the more open vocal tract. Especially the first formant frequency (F_1) increases and the bandwidth decreases while the second formant frequency (F_2) may decrease. The spectral emphasis of speech is also shifted from low frequencies in low vocal effort to mid or high frequencies in high vocal effort. High vocal effect is characterized by decreased spectral tilt, which is due to the increased subglottal pressure and increased vocal fold tension, thus creating a more abrupt closure of the vocal folds. Finally, one of the most prominent consequences caused by increased vocal effort is the rising of the sound pressure level (SPL) of speech, increasing the SNR and thus intelligibility of speech. However, it is important to note that the vocal effort is a subjective phenomenon, different from the purely objective quantitative measure represented by SPL. The effects of increased or decreased vocal effort are generally similar, but, as Junqua (1993) emphasizes, the effects of vocal effort may vary according to speaker. As reported by Summers et al. (1988), the effects of increased vocal effort are also related to those of clear speech.

Increased vocal effort, such as Lombard speech and shouting, has been studied relatively more than decreased vocal effort. Decreased vocal effort, such as a breathy and whispery voice, can be used, e.g., in the expression of emotions or in specific contexts. Although vocal effort is generally adjusted according to communication distance, Traunmüller and Eriksson (2000) discovered that speakers did not change their vocal effort substantially within the common communication distances (from 0.3 m to 1.5 m) in everyday conversations. Thus, decreased vocal effort may be less related to distance and more to the context. Breathy phonation is characterized by an increased non-harmonic noise component (aspiration) and an emphasized fundamental frequency component (increased spectral tilt). This is due to the relaxation of the laryngeal muscles, which leads to an incomplete closure of the vocal folds during vibration (Ishi et al., 2010). Whispers and whispery speech, in which turbulent noise is the main component of the voice (Ishi et al., 2010), are at the extreme end of the vocal effort continuum. Interestingly, Traunmüller and Eriksson (2000) observed that in whispered speech, formant frequencies and durations were affected similarly as with increased vocal effort. However, whispered speech is not explicitly considered in this study.

2.2. Modeling Vocal Effort

Vocal effort affects many properties of speech, as is described in the previous section. In speech recognition, changes in vocal effort might cause a mismatch between training and testing thereby resulting in decreased recognition accuracy (Junqua, 1993). Therefore, the effects of vocal effort have been investigated, at least to some extent, in speech recognition (e.g. Junqua, 1993; Zelinka et al., 2012), but the topic has remained less studied in speech synthesis. Although expressive speech synthesis has been an important goal since the development of the first synthesizers, there are few studies in which the vocal effort is explicitly modeled.

In concatenative speech synthesis (Hunt and Black, 1996; Black and Campbell, 1995), only few attempts have been made to reproduce vocal effort. In (Schröder and Grice, 2003; Turk et al., 2003), a diphone database of varying degree of vocal effort (soft, modal, and loud) was created and used to create varying degrees of vocal effort in synthesis. In (Langner and Black, 2005), voice conversion techniques were used to convert synthesis output to represent high vocal effort. In (Cernak, 2006; Patel et al., 2006), unit selection synthesis was modified by changing prosody generation and unit selection process in order to create synthetic speech that corresponds to speech-in-noise, thus increasing intelligibility. Indirect attempts to create a varying degree of vocal effort in unit selection synthesis can be found in the studies on expressive speech synthesis, but usually these concentrate more on the prosodic level of expression rather than on the acoustic level. Nevertheless, expressive concatenative speech synthesis is laborious due to the requirement of huge amount of speech data; a large corpus, from 3 to 10 hours, may be required for each different expressive style. Especially for high vocal effort speech, such a corpus with consistent quality is hard to achieve. Thus, unit selection approach is not very suitable for generating emotional expressions or different speaking styles (Black, 2003).

Varying the vocal effort of speech has been studied also in the field of voice conversion. For example, Huang et al. (2010) studied mimicking of Lombard speech by using the speech manipulation tool STRAIGHT (Kawahara et al., 1999, 2001), and Nordstrom et al. (2008) investigated the transformation of the singing voice from high vocal effort to breathy using adaptive pre-emphasis linear prediction. Despite the progress achieved in these studies, a comprehensive transformation of the vocal effort, taking into account all the varying characteristics, has proven to be difficult.

Statistical parametric speech synthesis (Zen et al., 2009), or hidden Markov model (HMM) based speech synthesis, offers an easier way to create variation in the synthetic voice. In HMM-based speech synthesis, the voice characteristics can be easily changed by transforming the HMM parameters. The most common transformation is the adaptation of the statistical speech models by parameters estimated from a small amount of speech (Yamagishi et al., 2009). Thus, only small amounts of vocal effort or expression specific speech is required for creating each desired voice type.

However, a major drawback of HMM-based speech synthesis is known to be its lower quality in terms of naturalness compared to unit selection synthesis. The most straightforward vocoder used in HMM-based synthesis, in which the excitation signal is composed of an impulse train in voiced and white noise in unvoiced sections, enables only low-quality synthesis with little flexibility in terms of expression. Only recent developments in the excitation and vocoder techniques (e.g. Kawahara et al., 1999, 2001; Drugman and Dutoit, 2012; Raitio et al., 2011c,a) have succeeded in providing the desired synthesis quality and a real ability for expressive speech synthesis.

While expressive HMM-based speech synthesis has been an active research topic for many years (e.g. Yamagishi et al., 2005; Tachibana et al., 2005, 2006; Nose et al., 2007) due to the suitability of the statistical paradigm for this purpose, only few studies can be found (e.g. Calzada and Socoró, 2011) that explicitly model vocal effort in HMM-based speech synthesis. Some interest in the research was raised by the Blizzard Challenge workshop in 2010 (King and Karaiskos, 2010), where one of the tasks was to create a voice that is maximally intelligible in the presence of noise. The GlottHMM system (Suni et al., 2010) that mimicked the Lombard effect by modifying the synthesis parameters was the most intelligible in the test, even being more intelligible than natural normal speech. The GlottHMM system has also been used to create Lombard speech with different adaptation techniques (Raitio et al., 2011b). The experiments in the paper indicated that the best approach was the over-adaptation (extrapolation) of the normal voice model with Lombard speech data. The synthetic Lombard speech was rated to be as intelligible and suitable in noisy conditions as natural Lombard speech.

3. Creating Vocal Effort Continuum

Previously in (Raitio et al., 2011b), only one male speaker was used to adapt normal speech to Lombard speech. For the current study, this setup was extended to incorporate the full vocal effort continuum, ranging from low effort

breathy speech to high vocal effort Lombard speech. In addition, two new speech corpora were applied, with both a female and a male speaker. The voices were constructed from three types of speech that was recorded along the continuum: breathy, normal, and Lombard speech. Corresponding synthetic voices were created by training and adapting the statistical parametric speech synthesis system GlottHMM. The process of creating these voices is described in this section.

3.1. Speech Material

Two new speech corpora were utilized consisting of one male and one female speaker, aged 47 and 49 years, respectively. Three different speaking styles were recorded from both subjects: breathy, normal, and Lombard style. Subjects were recorded in a soundproof studio, standing 10 cm away from an AKG 4000B large diaphragm condenser microphone. A secondary microphone and headphones were used for speaker feedback in Lombard and silent speech. Recordings were performed with a sampling rate of 48 kHz with 24-bit resolution.

For the normal style, both speakers read the same material, comprising approximately two hours of speech. The material consists of 500 phonetically rich isolated sentences, 230 long utterances of continuous non-fiction, 620 utterances of continuous prose, and 100 sentences of prosodically diverse prose quotations, as well as various shorter texts, such as names, numbers and lists. The Lombard speech was elicited from subjects by playing babble noise from the NOISEX-92 database with 80 dB SPL to the speaker's ears through Sennheiser HD 250 linear II headphones. The SPL of the noise was calibrated with a Cortex MK2 artificial head. The speaker's own feedback of speech was set by one of the authors to correspond to a situation of hearing his own voice in a quiet room without headphones. The feedback level was kept constant for both subjects. For the Lombard style, the subjects read a subset of the material, 200 phonetically rich sentences and 100 sentences of prosodically rich quotations. The breathy speaking style was elicited by increasing the level of the speaker's own feedback through headphones as well as instructing the subjects to try to keep their speech voiced, i.e., not to whisper. One hundred phonetically rich sentences and 100 quotations were read in the breathy style.

The speech material was manually checked, corrected, and split into utterances. Further annotation was performed on the material during voice training. Segmentation was performed, and phrase breaks were annotated by force-alignment of speaker- and style-dependent monophone models. Word prominences were labeled by comparing the generated prosodic parameters of simple context-dependent HMMs with the original parameters (Suni et al., 2012), analyzed with GlottHMM.

A prosodic analysis of the resulting corpora was carried out using 100 isolated sentences of each style. Both speakers' raw f_0 values were analyzed; the resulting distributions are shown in Figures 1 and 2 for the female and male speaker, respectively. The distributions are similar for both speakers with regard to the speaking style; the breathy style results in a relatively low average f_0 with a very narrow distribution, whereas the Lombard speech has a wider distribution with a relatively high mean. The normal speech is between the two extremes, but with a large overlap with the breathy style. The distributions of logarithmic segmental durations for both speakers were also studied. The female speaker has an overall slower speaking rate with an average duration of 80.4 ms as opposed to the male with an average duration of 69.05 ms. With respect to statistical significance, the male speaker's distributions did not differ from each other between different speaking styles, whereas the female speaker's Lombard speech was significantly slower than the other styles (t-tests, $p < 0.0001$), which in turn were not significantly different.

3.2. Speech Synthesis System

The statistical parametric speech synthesis system GlottHMM (Raitio et al., 2011c) was used to generate the synthetic speech in this study. GlottHMM is built on a basic framework of a HMM-based speech synthesis system (Zen et al., 2007), but it uses a distinct type of vocoder for parametrizing and synthesizing speech. GlottHMM aims to accurately model the speech production mechanism by decomposing speech into the vocal tract filter and the voice source signal using glottal inverse filtering and emphasizing the modeling of the voice source.

GlottHMM has been constantly developed since its conception (Raitio et al., 2008). It is thoroughly described in (Raitio et al., 2011c), but further developments have been made to the system since then. Some of the developments are described in (Suni et al., 2010, 2011, 2012), but in order to give an up-to-date and concise description, the GlottHMM system configuration used in this study is described next.

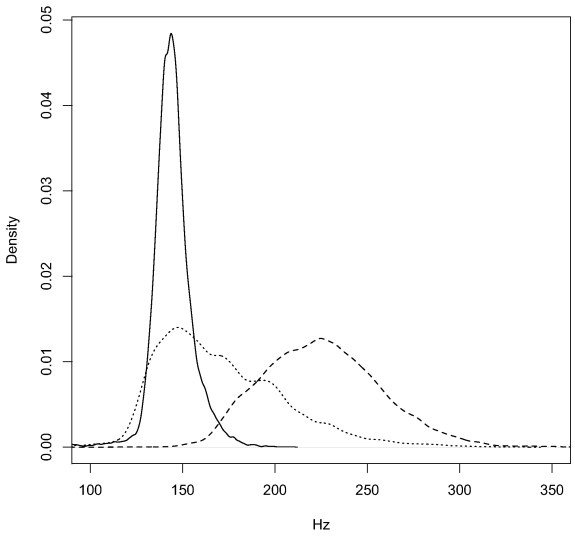

Figure 1: Female f_0 distributions of the breathy (—), normal (···), and Lombard (---) speech.

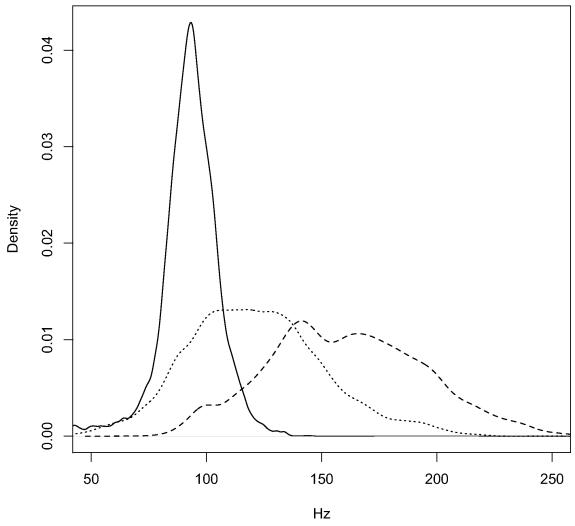

Figure 2: Male f_0 distributions of the breathy (—), normal (···), and Lombard (---) speech.

3.2.1. Parametrization

The flow chart of the parametrization stage of GlottHMM is shown in Figure 3. First, the speech signal is high-pass filtered with a cut-off frequency of 70 Hz in order to remove possible low-frequency fluctuations that may distort the glottal flow estimate. The speech signal is then windowed into two types of frames: a short frame is used to extract the vocal tract filter, the voice source spectral envelope, and the short time energy of speech. A longer frame is used for estimating several pitch periods of the glottal flow, which many of the voice source parameter require. The length of the shorter frame is 25 ms, whereas the length of the longer frame depends on the f_0 range of the speaker.

For glottal inverse filtering, iterative adaptive inverse filtering (IAIF) (Alku, 1992; Alku et al., 1999) is used. The modified version of the method (Suni et al., 2011) is used in order to make the glottal flow signal estimation more robust at the cost of losing some accuracy. However, stable parameter values are required for the training of the parameters in HMMs. The algorithm uses linear prediction (LP) for estimating the spectral envelope of speech. The outputs of the modified IAIF algorithm are the estimated vocal tract LP filter and the estimated glottal flow signal. The vocal tract filter is converted into line spectral frequencies (LSFs), a parametric representation of LP information well-suited for use in a statistical parametric speech synthesis system (Marume et al., 2006), providing stability (Soong and Juang, 1984) and low spectral distortion (Paliwal and Kleijn, 1995).

The longer frame is processed with the same modified IAIF algorithm to estimate the glottal flow signal. Ideally, at least two complete pitch periods are included in the glottal flow estimate even with the lowest f_0 values. The glottal flow signal is used to define f_0, estimated with the autocorrelation method. Harmonic-to-noise ratio (HNR) of the signal indicates the degree of voicing, i.e., the relative amplitudes of the periodic vibratory glottal excitation and the aperiodic noise component of the excitation. The HNR is based on the ratio between the upper and lower smoothed spectral envelopes (defined by the harmonic peaks and interharmonic valleys, respectively) and it is averaged across five frequency bands according to the equivalent rectangular bandwidth (ERB) scale (Moore and Glasberg, 1996). Finally, the glottal closure instants (GCIs) are detected by a simple peak-picking algorithm that searches for the negative excitation peaks of the glottal flow derivative at fundamental period intervals. For all the two-pitch period speech segments found, the modified IAIF algorithm is applied pitch-synchronously again in order to yield a better estimate of the glottal flow. The re-estimated two-period glottal flow derivative waveforms are windowed with the Hann window, and a pulse library is constructed from the extracted waveforms. The speech features extracted by the vocoder are depicted in Table 1.

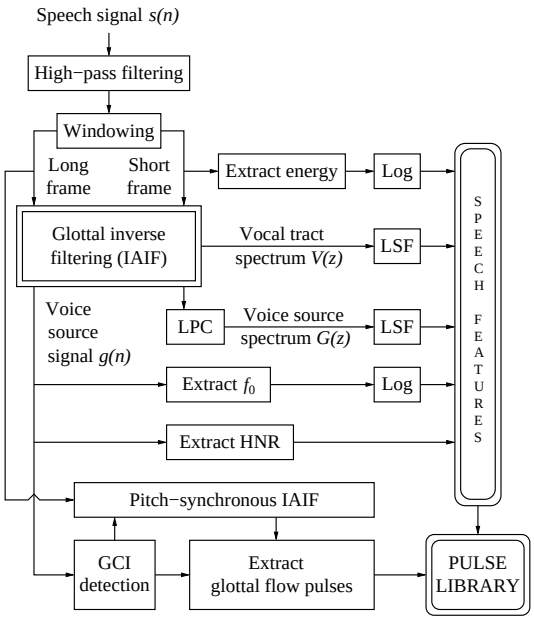

Figure 3: Flow chart of the parametrization stage.

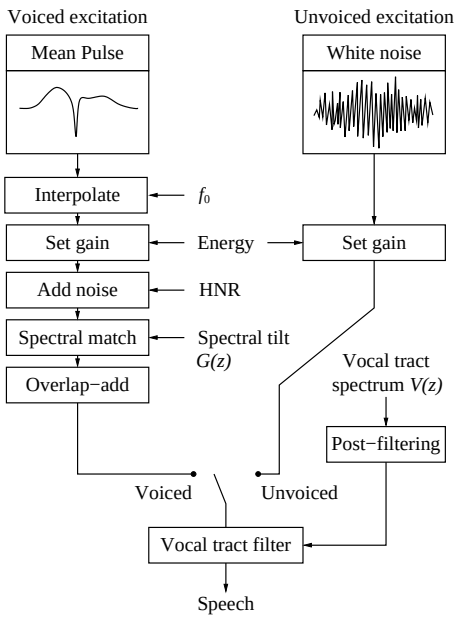

Figure 4: Flow chart of the synthesis stage.

3.2.2. *Synthesis*

There are several ways to utilize the glottal-flow-pulse library that is obtained at the end of the parametrization, as described in Section 3.2.1. In the original implementation of GlottHMM (Raitio et al., 2008, 2011c), only a single pulse was used and modified to create the voiced excitation. In a more complex framework (Raitio et al., 2011a), a pulse from the library is selected for each time instant by minimizing the target cost of the voice source parameters and concatenation cost of the pulse waveforms, and finally the selected pulses are concatenated to create the excitation signal. In this study, an approach similar to the one in (Drugman and Dutoit, 2012) was adopted by using only the mean of the pulse library as a basis for the voiced excitation. Although the use of the pulse library method may yield higher quality with some voices, the simpler method probably yields more consistent behavior between the different speaking styles.

The flow chart of the synthesis stage is illustrated in Figure 4. The basis of voiced excitation is the mean of the extracted two-period glottal flow derivative waveforms in the library that are interpolated to a constant length. This pulse signal is interpolated in the time domain with a cubic spline interpolation algorithm (Engeln-Müllges and Uhlig, 1996; Galassi et al., 2009) in order to achieve a specific fundamental period and the amplitude of the pulse is scaled based on the energy measure. In order to control the degree of voicing, the amount of noise in the excitation is matched by manipulating the phase and magnitude of the spectrum of each pulse based on the HNR at each ERB band. Furthermore, the spectral tilt of each pulse is modified depending on the all-pole spectrum generated by the HMM. This modification is achieved by filtering the pulse train with an adaptive infinite impulse response (IIR) filter,

Table 1: Speech features and the number of parameters.

Feature	Number of parameters	Type/Unit	Contribution
Vocal tract spectrum	30	LSF	Vocal tract filter
Energy	1	dB	Voice source
Fundamental frequency	1	$\log f_0$	Voice source
Harmonic-to-noise ratio	5	dB/ERB	Voice source
Voice source spectrum	10	LSF	Voice source
Pulse library	1200–3600	Pulse waveform	Voice source

which flattens the spectrum of the pulse train and applies the desired spectrum. The IIR filter is constructed from the generated voice source LP spectrum (denominator) and the LP spectrum of the pulse (numerator). The unvoiced excitation is composed of white noise, whose gain depends on the energy measure generated by the HMM system.

Formant enhancement based on all-pole modeling (Raitio et al., 2010) is applied to the vocal tract LSFs in order to reduce over-smoothing caused by the statistical modeling. Finally, the LSFs are reconverted to LP coefficients describing the vocal tract spectrum and used for filtering the combined excitation signal.

3.3. Building Synthetic Voices

In constructing the synthetic voices, speech signals were first downsampled to 16 kHz and then parametrized with the GlottHMM vocoder according to Table 1. Full-context labels were generated with conventional features: quinphones with positional features of various units in utterance hierarchy. Additionally, content-function class and prominence labels for words were used.

The HMMs consisted of four streams: (1) the vocal tract spectrum LSFs combined with energy, (2) the voice source spectrum LSFs, (3) the harmonic-to-noise ratio (HNR), and (4) the fundamental frequency, f_0. The number of static parameters for each stream corresponds to the number of parameters given in Table 1, with the exception that the vocal tract LSFs and energy are combined. With delta and delta-delta features added, a total of 141 acoustic parameters were used in the HMM training. The training procedure of the normal voices followed the standard HTS method (Zen et al., 2007) with individual clustering of the streams and two iterations of clustering and re-estimation stages for full-context models.

However, some problems were encountered with this setup for the female voice. Examining the forced alignment of the full context models revealed that the embedded re-estimation had failed to find a correct path through some of the longer utterances. Thus, the novel idea of using typical speech recognition features in training was experimented with. A stream of 13 mel-cepstral coefficients, including the zeroth coefficient, and delta and delta-delta features was added to guide the training process in order to improve the alignments. Here, stream weights of the mel-cepstrum stream and f_0 were set to one and others to zero. This addition seemed to improve the modeling in this case, but further experiments are required to validate this approach.

In creating the low and high vocal effort models (breathy and Lombard), the normal voice models were adapted with constrained structural a maximum a posteriori linear regression combined with maximum a posteriori (CSMAPLR + MAP) adaptation technique (Yamagishi et al., 2009). Adaptation was applied to all streams using pruned state-tying decision trees for regression classes. The amount of adaptation data was the same for both speakers, 300 sentences in Lombard style and 200 sentences in breathy style.

For the evaluation, 80 utterances were synthesized with each resulting voice. Full-context labels of the utterances were generated with phrasing, and word prominence was predicted by rule. Speech parameter trajectories were generated taking into account the global variance (GV) of the original training data (Toda and Tokuda, 2007). The strength of the GV adjustment for each stream and the strength of formant enhancement were optimized manually in order to find a suitable combination for all styles. No individual tuning was performed on individual voices. The Lombard effect of the male voice was found to be substantially weaker than the female counterpart (due to personal variation in the Lombard effect or inability to properly elicit Lombard speech), and the mismatch was reduced by using an extrapolation ratio of 1.2 for the male voice, where 0.0 would be the normal voice and 1.0 the adapted Lombard voice. The extrapolation ratio was selected experimentally in order to match the voice with the recorded sentences that exhibited the strongest Lombard effect. Previous experiments (Raitio et al., 2011b) suggest that moderate extrapolation can be used without causing degradation in speech quality. Although in this experiment only three voices were created that approximately correspond to the three recorded corpora, it is important to note that the method enables creating voices in arbitrary positions on the continuum.

For all voice types, five sentences of the style-specific speech were used for extracting the pulse library. The number of pulses for calculating the mean was 1566, 1901, and 3562 for female and 1279, 1488, and 1876 for male breathy, normal, and Lombard speech, respectively. The corresponding mean pulses for different vocal effort levels for the female and male speaker are shown in Figure 5.

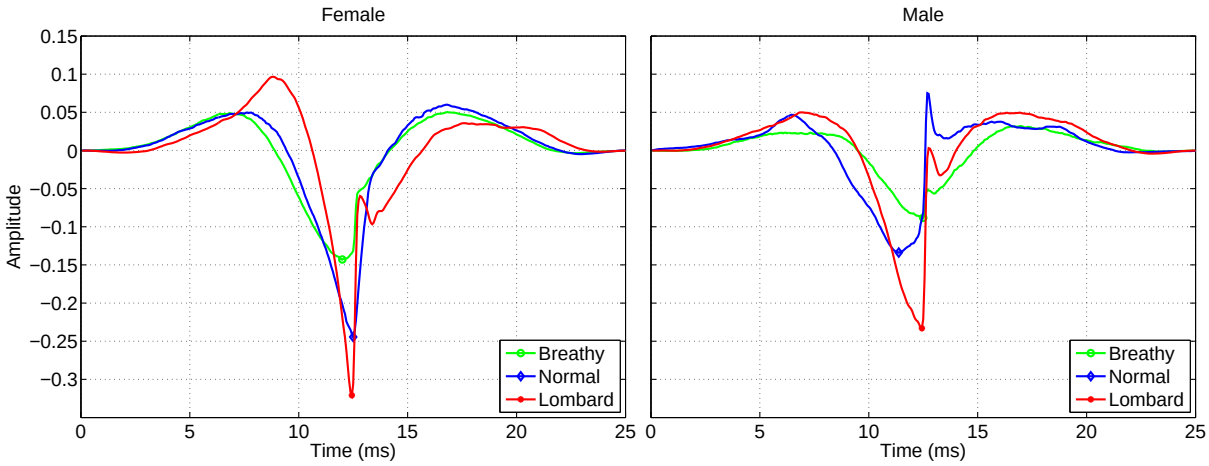

Figure 5: Illustration of the mean of the windowed two-period glottal flow derivative waveforms (pulses) for different vocal effort levels for the female (left) and male (right) speaker.

4. Evaluation

The aim of the evaluation was to assess various perceived characteristics of natural and synthetic speech at three points on the vocal effort continuum in different noise conditions. The intelligibility of speech was one of the main measures, but the importance of speech quality in silence and in noise was also addressed. Finally, the question of contextual appropriateness of speech with respect to the noise environment was considered. Additionally, the overall pleasantness of the voices was queried in order to find out whether the subjects would prefer to listen to low-effort speech, especially in silent conditions.

4.1. Noise Conditions

In order to simulate natural human speech communication, a realistic noise environment was created (Raitio et al., 2011b, 2012). Multichannel recordings of street noise was used, containing mostly traffic noise with most of the energy in the low frequencies. A 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 (Pulkki, 2007).

Three different environments were selected for the test: silence, moderate street noise, and extreme street noise. The averaged A-weighted SPLs were 63 dB and 70 dB for the moderate and extreme street noise cases, respectively. These noise levels were selected based on pre-listening of the speech samples with different noise levels. In this procedure, levels which resulted in the most prominent differences in intelligibility between the test cases were chosen. Thus, the average SNRs were -1 dB and -8 dB for moderate and extreme noises, respectively. The noise levels were the same as used in the experiments in (Raitio et al., 2011b, 2012).

4.2. Listening Environment

Listening tests were performed in a standardized listening room that is in accordance with the recommendation ITU-R BS.1116-1 (ITU, 1997). The listening room contains nine identical loudspeakers (Genelec 8260A) positioned at 2.4–2.6-meter distances around the listener. The speaker setup is illustrated in Figures 6 and 7. The loudspeaker responses at the listener position were equalized using DSP implemented in the loudspeakers by the manufacturer. The loudspeaker directly in the front was used to play the speech samples, and all of the nine loudspeakers were employed in the reproduction of the masking noise.

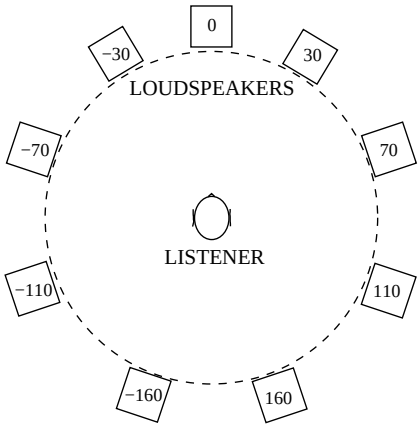

Figure 6: Illustration of the loudspeaker setup. Nine identical loudspeakers were positioned around the listener (angles depicted for each loudspeaker) in order to create a realistic noise environment.

Figure 7: A photograph of the listening test setup. The two rear loudspeakers at angles ±160° are not visible in the photo. The ceiling loudspeakers were not used in the test.

4.3. Speech Signals

Three levels of vocal effort were used in the test: breathy, normal, and Lombard. Both natural and synthetic speech signals were utilized for each vocal effort type for one female and one male speaker. Thus, the total number of speech signal types was six per gender. The active speech level of all speech signals was normalized according to ITU-T P.56 (ITU, 2011). After the normalization, the averaged A-weighted SPL of each voice type was measured in the listening environment through the test setup. All the six voice types, both female and male, and their measured SPLs are shown in Table 2. Interestingly, the SPLs show a wider range in different styles for the female speaker compared to the male, and informal perceptual observations of the corpora supported this finding.

The average spectra of female and male natural speech spoken in the three vocal effort levels are shown in Figure 8. The overall spectra and spectra for only voiced and unvoiced speech are shown separately. The average spectrum of the masking noise is also shown, but the level of the noise compared to speech is arbitrary. Inspecting the figures reveals that the Lombard voices are generally higher in magnitude in the low to mid frequencies compared to voices with less effort, but they also exhibit less energy in the high frequencies. This effect of increased vocal effort is due to two reasons. First, the decreased spectral tilt increases the magnitude of mid frequencies, and, second, the voiced component is relatively stronger compared to the unvoiced component. Since the voices are normalized in energy that is mostly determined by the voiced component, the unvoiced component becomes stronger in the normalized normal and breathy voices. This reflects the change in the balance between the voiced and unvoiced energy when the vocal effort is increased. Interestingly, the increase in the energy of voiced speech along with the increase of vocal effort is more noticeable in female speech that in male. On the other hand, the increase in the energy of the unvoiced component along with decreased vocal effort is more noticeable in male speech than in female.

The average spectral differences between natural and synthetic speech in the three vocal effort levels are shown in Figure 9. Natural and synthetic spectra are generally very similar, especially in the low frequencies. Interestingly, the high frequency spectra of synthetic speech varies most in the case of normal speech, especially with the male speaker

Table 2: Test voice types and their averaged A-weighted SPLs after loudness normalization with ITU-T P.56.

Type	Description	SPL (Female)	SPL (Male)
nat_bre	Natural breathy speech	58 dB	60 dB
syn_bre	Synthetic breathy speech	58 dB	60 dB
nat_nor	Natural normal speech	63 dB	61 dB
syn_nor	Synthetic normal speech	62 dB	61 dB
nat_lom	Natural Lombard speech	65 dB	64 dB
syn_lom	Synthetic Lombard speech	65 dB	64 dB

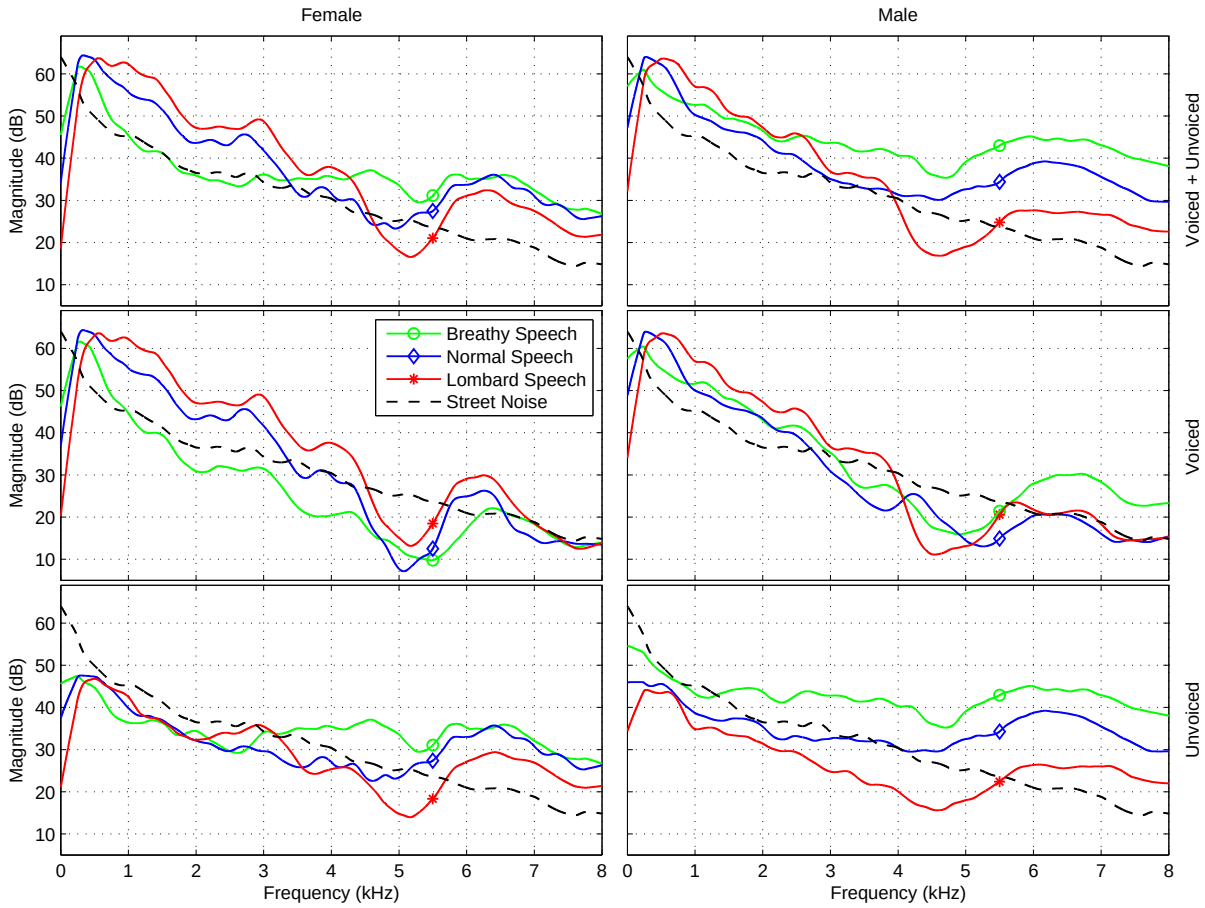

Figure 8: Average spectra of natural breathy, normal, and Lombard speech for female (left) and male (right) speakers. The uppermost graphs represent the overall spectrum, the middle graphs represent only voiced spectra, and the lowest only unvoiced spectra. The spectrum of the masking street noise is shown in the dashed black line (the level of the noise is arbitrary compared to speech spectra).

where high frequencies are much lower in magnitude compared to natural speech. The spectral dip between 7.5 kHz and 8.0 kHz is due to the statistical modeling of the vocal tract spectrum with LSFs, but it has presumably a very small perceptual effect.

Instead of using phonetically complex or semantically unpredictable sentences, ordinary short sentences were involved in the test. The sentence sets were designed to have a closely matching distribution of normal language with regards to phoneme and lexical frequencies, word length, etc. (Vainio et al., 2005). Due to the lack of large amounts of natural speech data, different sentences from the same database were used to test natural and synthetic speech. For natural speech, the test set consisted of 48 recorded utterances, and for synthetic speech, 80 sentences were used. For each test case, a random set of 36 sentences from both natural and synthetic sets was selected. A representative set of the test voices is available online at http://www.helsinki.fi/speechsciences/synthesis/samples.html.

4.4. Listening Tests

Two types of assessments were performed in order to evaluate the characteristics of the voices. First, an intelligibility test was performed, in which short Finnish sentences were presented to the listener either in silence or masked by the noise (63 dB or 70 dB). The listener was allowed to listen to the sample only once. The task of the listener was to type in what she or he heard. Word error rates (WERs) of the answers were then evaluated.

Second, the speech samples were subjectively rated according to three questions with continuous scales with five verbal attributes:

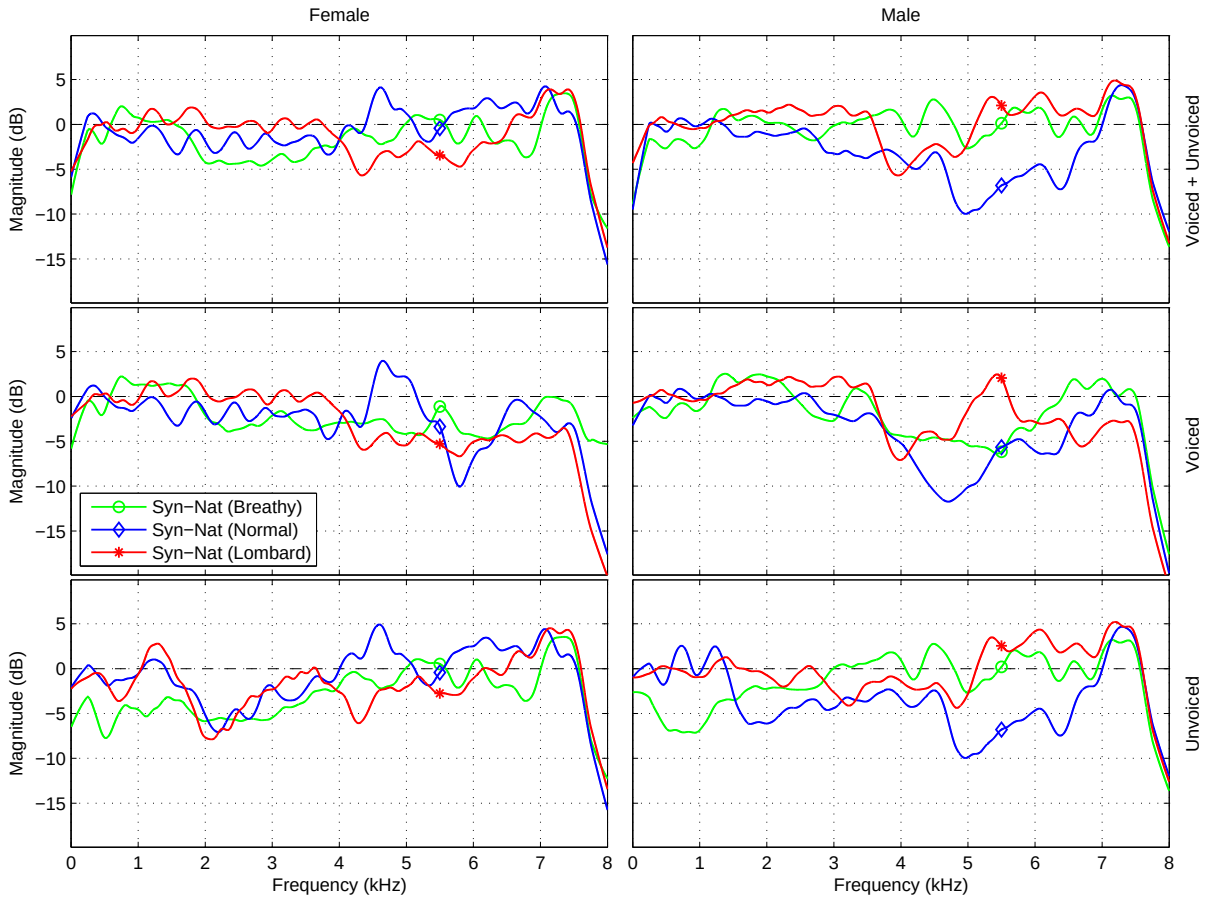

Figure 9: Average spectral differences between natural and synthetic speech of breathy, normal, and Lombard style for female (left) and male (right) speakers. The uppermost graphs represent the overall spectrum, while the middle and the lowest graphs represent only voiced and only unvoiced spectra, respectively.

1. How would you rate the quality of the speech sample?
   (*bad – poor – fair – good – excellent*)
2. How suitable was the speaking style considering the sound environment?
   (*bad – poor – fair – good – excellent*)
3. How would you describe the speaking style?
   (*very irritating – slightly irritating – neutral – slightly pleasant – very pleasant*)

The continuous scale ratings, anchored by the five equally spaced verbal attributes, were then scaled to correspond to values from 0 to 100.

Every listener rated two sentences of each of the six speech types in all three noise conditions for both genders. Thus each listener rated a total of 72 speech samples in both test types (intelligibility and subjective rating). The test took approximately one hour per listener. A total of 27 listeners (5 females and 22 males) with no reported hearing disorders took part in the test. The listeners were young university students, and they were paid for participating in the test.

4.5. Results of the Intelligibility Test

The results of the intelligibility test in all noise conditions for female and male voices are shown in Figure 10. The WERs are shown for each case with 95% confidence intervals. In silence, all female voice types are equally

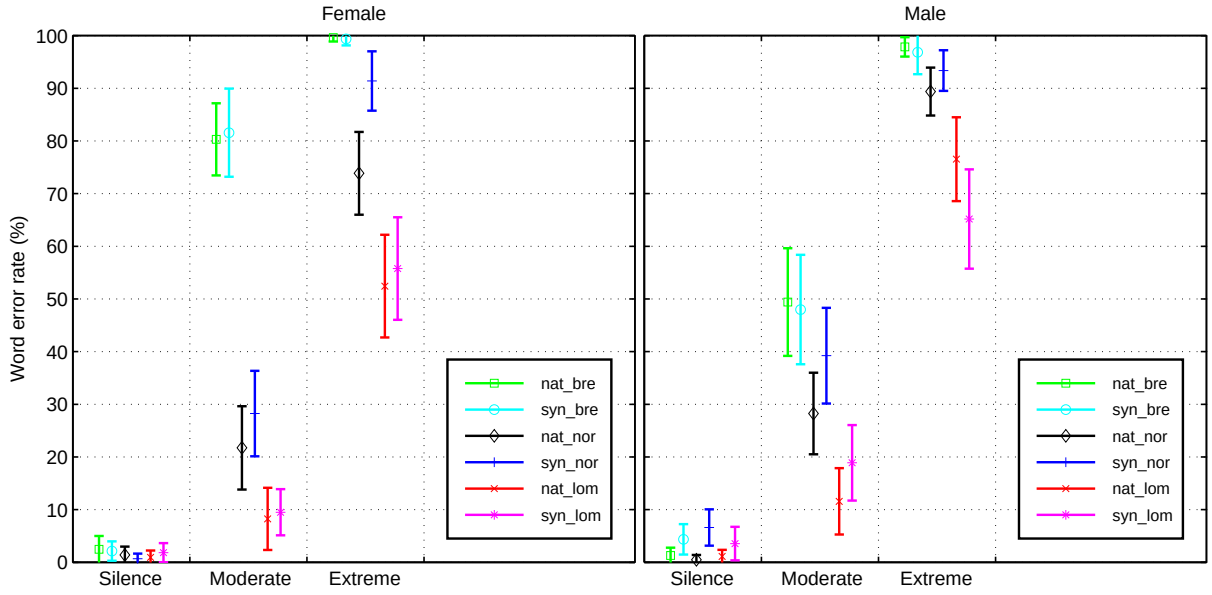

Figure 10: Results of the intelligibility test for female (left) and male (right) voices in three noise conditions: silence, moderate street noise (63 dB, SNR = −1 dB), and extreme street noise (70 dB, SNR = −8 dB).

intelligible with WERs of 1–3%, but for the male voice, the synthetic ones tend to have a slightly larger WER than the natural ones. Informally, the male voice was observed to have some errors in certain phoneme combinations, indicating slight errors in the HMM training.

The differences in intelligibility with different vocal effort level start to show under the moderate noise condition. Female breathy voices are almost totally unintelligible with WERs around 80%, while normal and Lombard voices have significantly lower WERs, as expected. Male breathy voices do not have as high WERs (around 50%) as the female breathy voices, but they are still generally less intelligible than normal and Lombard voices. The synthetic voices in moderate noise tend to have slightly greater WERs than the natural ones, but the difference is not very large.

In extreme noise, both female and male breathy voices are almost totally unintelligible with WERs of 97–100%. Normal voices are slightly more intelligible with WERs of 74–93% whereas Lombard voices show the smallest WERs of 52–77%. The female Lombard voice is clearly more intelligible than the male Lombard voice. In (Raitio et al., 2011b), another low-pitched male voice with a strong Lombard effect was evaluated in an identical test setup with very low WERs, even lower than the female voice in this study. Thus the low intelligibility of the male voice in this study is most probably due to the relatively weak personal Lombard effect of the speaker (see Table 2). Interestingly, the synthetic male Lombard voice was more intelligible than the natural male Lombard voice. This is probably due to the effect of tuning the extrapolation coefficient in the adaptation, and in this case, the synthetic voice is slightly more "Lombard" than the original Lombard voice.

A further analysis of the results of the intelligibility test revealed that most of the misrecognized words were mono or disyllabic function words (including copula), which are usually reduced both prosodically and segmentally. On the other hand, more uncommon words relative to the test corpus and in general, which are difficult to predict from the context, were also often misrecognized.

4.6. Results of the Subjective Evaluation

The results of the subjective evaluations in all noise conditions for female and male voices are shown in Figure 11. The subjective ratings according to the questions and continuous scales with five verbal descriptions are represented as values from 0 to 100 with 95% confidence intervals. The quality ratings show that the natural voices are rated higher than the synthetic voices in silence. The difference in quality ratings between natural and synthetic voices is smaller but still notable in moderated noise, but in extreme noise the differences almost completely disappear.

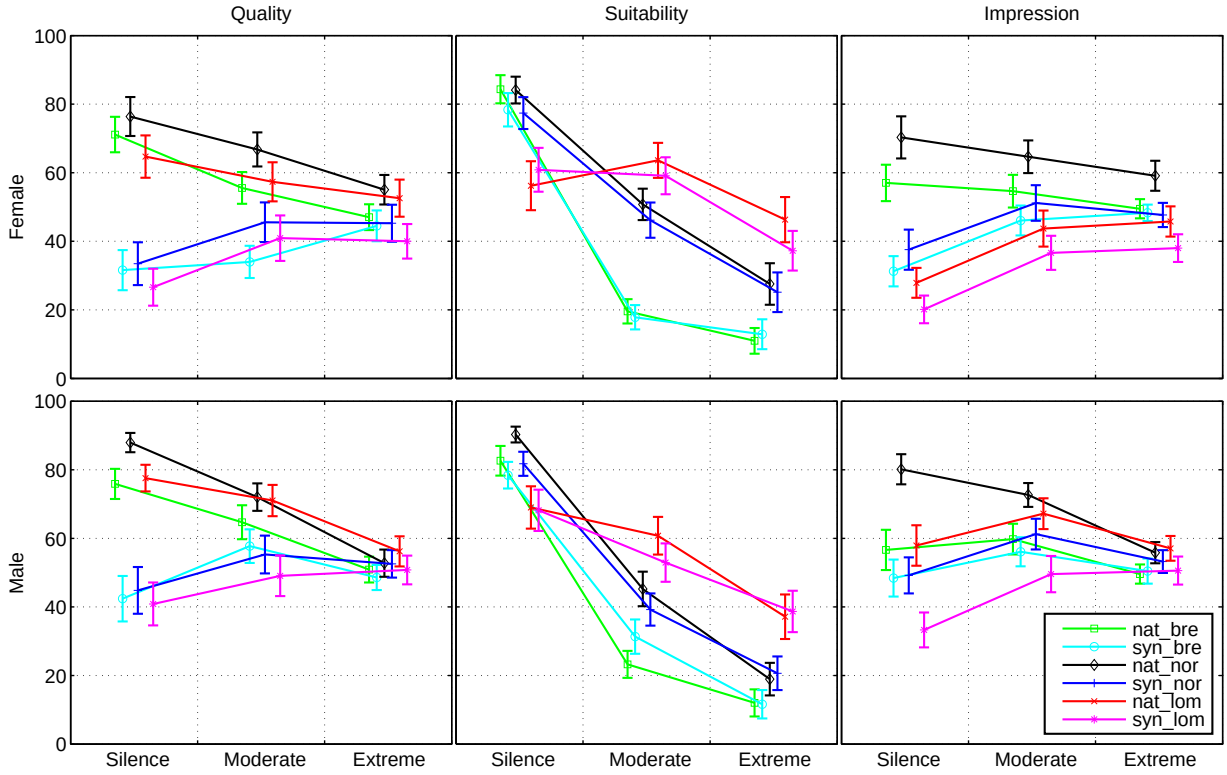

Figure 11: Results of the subjective evaluation for female (upper) and male (lower) voices. The measured quantities are quality, suitability, and impression (see questions and scales in Section 4.4).

The suitability ratings of the voices in different sound environments show that breathy and normal speech are rated as very suitable in silence, whereas Lombard speech is rated as slightly less suitable, although still suitable rather than not suitable. In moderate noise, the ratings are reversed. Normal and especially breathy voices are rated low in suitability, indicating that the level of vocal effort was not adequate considering the noise. Lombard voices are rather suitable (fairly or well suitable) in moderate noise. No major differences can be found between natural and synthetic voices with regard to suitability, indicating that the reproduction of vocal effort was successful.

The impression rating measures whether the speaking style was irritating, neutral, or pleasant. The results show no clear evidence that decreased vocal effort would increase pleasantness. For both female and male natural voices in silent conditions, normal style was rated as more pleasant than breathy style. For Lombard style, the results diverged. The female voice was rated substantially lower than the male counterpart. Increasing the background noise diminishes the differences, but, unlike in the case of suitability, the normal styles are rated the most pleasant. Generally, natural voices are rated higher than synthetic ones, as expected. Lombard style synthesis is considered somewhat less pleasant than synthetic normal and breathy styles. No significant differences between synthetic breathy and normal style were found.

5. Conclusions

This study presented a method to create synthetic voices over a wide vocal effort continuum. The method consists of first recording databases of breathy, normal, and Lombard speech and then using statistical parametric speech synthesis and adaptation methods to create desired synthetic voices. The method enables speech synthesis at any arbitrary point on the continuum, even slightly outside the two end points (breathy or Lombard).

Natural and synthetic speech along the continuum were assessed in listening tests that evaluated the intelligibility, quality, and suitability in three different realistic multichannel noise conditions: silence, moderate street noise, and

extreme street noise. The results of the evaluation show that increased vocal effort improves the intelligibility of speech both for natural and synthetic voices. Although the synthetic voices have generally slightly higher WERs than natural speech, the reproduction of vocal effort in synthesis was successful. In the case of synthetic male Lombard speech, the WER was even lower than with natural Lombard speech. This result is most likely due to the different voice building method; in male Lombard voice, extrapolation was used to achieve a voice that has enough Lombard characteristics. A larger extrapolation coefficient in adaptation caused the synthetic Lombard voice to have a slightly higher level of vocal effort than the original Lombard voice.

The results of the subjective evaluation show that the synthesized voices with varying vocal effort are rated very similarly to their natural counterparts. Especially the suitability ratings of natural and synthetic voices have a very high correlation. Only the quality ratings show a strong separation between the natural and synthetic voices. The Lombard voices are generally more suitable to be heard in the presence of noise compared to conventional voices and they are rated very similar to natural Lombard speech, as was also reported in (Raitio et al., 2011b). However, the breathy voice was not considered more appropriate nor more pleasant in silence than normal speech. This result is consistent with the results of Traunmüller and Eriksson (2000). They reported that speakers did not change their vocal effort substantially within the common communication distances in everyday conversations. Thus, breathy voice is probably a more socially dependent speaking style. On the other hand, breathy voice is also related to expression of emotions. These contexts were not intended to be reproduced in the current experiment and thus breathy voice was not rated more appropriate than normal voice. Nevertheless, breathy voice was rated as very suitable in silence.

The study also showed that the quality of speech does not seem to suffer significantly from adaptation; all the synthetic voices with different vocal effort levels were rated equal in quality in silence. Also an important outcome of the evaluation was that, in the presence of noise, the degradation of speech quality caused by statistical modeling and vocoding loses its significance, thus justifying the use of such synthetic voices in the presence of noise.

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 the Academy of Finland (projects 135003 LASTU programme, 1128204, 1218259, 121252).

References

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

Alku, P., Tiitinen, H., Näätänen, R., 1999. A method for generating natural-sounding speech stimuli for cognitive brain research. Clinical Neurophysiology 110, 1329–1333.

Black, A., 2003. Unit selection and emotional speech. In: Proc. Eurospeech '03. pp. 1649–1652.

Black, A., Campbell, N., 1995. Optimising selection of units from speech database for concatenative synthesis. In: Proc. Eurospeech '95. pp. 581–584.

Calzada, A., Socoró, J., 2011. Vocal effort modification through harmonics plus noise model representation. In: Travieso-González, C. M., Alonso-Hernández, J. B. (Eds.), Advances in Nonlinear Speech Processing. Vol. 7015 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 96–103.

Cernak, M., 2006. Unit selection speech synthesis in noise. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 14–19.

Drugman, T., Dutoit, T., 2012. The deterministic plus stochastic model of the residual signal and its applications. IEEE Trans. on Audio, Speech, and Lang. Proc. 20 (3), 968–981.

Engeln-Müllges, G., Uhlig, E., 1996. Numerical Algorithms with C. Springer, Berlin.

Galassi, M., et al., 2009. GNU Scientific Library Reference Manual. 3rd Edition.

Gobl, C., Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Commun. 40 (1-2), 189–212.

Gordon, M., Ladefoged, P., 2001. Phonation types: a cross-linguistic overview. Journal of Phonetics 29 (4), 383–406.

Huang, D.-Y., Rahardja, S., Ong, E. P., 2010. Lombard effect mimicking. In: Seventh ISCA Workshop on Speech Synthesis. pp. 258–263.

Hunt, A., Black, A., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 373–376.

Ishi, C. T., Ishiguro, H., Hagita, N., 2010. Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. EURASIP J. Audio Speech Music Process. 3, 1–12.

ITU, 1997. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. International Telecommunication Union, Recommendation ITU-R BS.1116-1.

ITU, 2011. Objective measurement of active speech level. International Telecommunication Union, Recommendation ITU-T P.56.

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

Kawahara, H., Estill, J., Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In: 2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA).

Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. 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 Commun. 27 (3–4), 187–207.

King, S., Karaiskos, V., 2010. The Blizzard Challenge 2010. In: The Blizzard Challenge 2010 workshop. http://festvox.org/blizzard.

Langner, B., Black, A. W., 2005. Improving the understandability of speech synthesis by modeling speech in noise. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 265–268.

Lombard, E., 1911. Le signe de l'elevation de la voix. Ann. Maladies Oreille, Larynx, Nez, Pharynx 37 (101–119), 25.

Marume, M., Zen, H., Nankaku, Y., Tokuda, K., Kitamura, T., 2006. An investigation of spectral parameters for HMM-based speech synthesis. In: Proc. Autumn Meeting of Acoust. Soc. of Japan. (In Japanese).

Moore, B., Glasberg, B., 1996. A revision of Zwicker's loudness model. ACTA Acustica 82, 335–345.

Nordstrom, K., Tzanetakis, G., Driessen, P., 2008. Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction. IEEE Trans. on Audio, Speech, and Lang. Proc. 16 (6), 1087–1096.

Nose, T., Yamagishi, J., Kobayashi, T., 2007. A style control technique for HMM-based expressive speech synthesis. IEICE Trans. Inf. Syst. E90-D (9), 1406–1413.

Paliwal, K., Kleijn, W., 1995. Quantization of LPC parameters. In: Kleijn, W., Paliwal, K. (Eds.), Speech Coding and Synthesis. Elsevier, Ch. 12.

Patel, R., Everett, M., Sadikov, E., 2006. Loudmouth: Modifying text-to-speech synthesis in noise. In: 8th intl. ACM SIGACCESS conf. on Computers and Accessibility.

Pulkki, V., 2007. Spatial sound reproduction with directional audio coding. J. Audio Eng. Soc. 55 (6), 503–516.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., Alku, P., 2008. HMM-based Finnish text-to-speech system utilizing glottal inverse filtering. In: Proc. Interspeech. pp. 1881–1884.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., Alku, P., 2010. Comparison of formant enhancement methods for HMM-based speech synthesis. In: Seventh ISCA Workshop on Speech Synthesis. pp. 334–339.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., Alku, P., 2011a. Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4564–4567.

Raitio, T., Suni, A., Vainio, M., Alku, P., 2011b. Analysis of HMM-based Lombard speech synthesis. In: Proc. Interspeech. pp. 2781–2784.

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

Raitio, T., Takanen, M., Santala, O., Suni, A., Vainio, M., Alku, P., 2012. On measuring the intelligibility of synthetic speech in noise – Do we need a realistic noise environment? In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4025–4028.

Rostolland, D., 1982. Acoustic features of shouted voice. Acustica 50, 118–125.

Schröder, M., Grice, M., 2003. Expressing vocal effort in concatenative synthesis. In: Proc. 15th International Conference of Phonetic Sciences. pp. 2589–2592.

Soong, F. K., Juang, B.-H., 1984. Line spectrum pair (LSP) and speech data compression. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vol. 9. pp. 37–40.

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

Suni, A., Raitio, T., Vainio, M., Alku, P., 2010. The GlottHMM speech synthesis entry for Blizzard Challenge 2010. In: The Blizzard Challenge 2010 workshop. http://festvox.org/blizzard.

Suni, A., Raitio, T., Vainio, M., Alku, P., 2011. The GlottHMM entry for Blizzard Challenge 2011: Utilizing source unit selection in HMM-based speech synthesis for improved excitation generation. In: The Blizzard Challenge 2011 workshop. http://festvox.org/blizzard.

Suni, A., Raitio, T., Vainio, M., Alku, P., 2012. The GlottHMM entry for Blizzard Challenge 2012 – Hybrid approach. In: The Blizzard Challenge 2012 workshop. http://festvox.org/blizzard.

Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T., 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Trans. Inf. Syst. E88-D (11), 2484–2491.

Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T., 2006. A style adaptation technique for speech synthesis using HSMM and suprasegmental features. IEICE Trans. Inf. Syst. E89-D (3), 1092–1099.

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

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

Turk, O., Schröder, M., Bozkurt, B., Arslan, L., 2003. Voice quality interpolation for emotional text-to-speech synthesis. In: Proc. Interspeech. pp. 797–800.

Vainio, M., Suni, A., Järveläinen, H., Järvikivi, J., Mattila, V.-V., 2005. Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish. J. Acoust. Soc. Am. 118 (3), 1742–1750.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009. 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.

Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., 2005. Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. IEICE Trans. Inf. Syst. E88-D (3), 503–509.

Zelinka, P., Sigmund, M., Schimmel, J., 2012. Impact of vocal effort variability on automatic speech recognition. Speech Commun. 54 (6), 732–742.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., Tokuda, K., 2007. The HMM-based speech synthesis system (HTS) version 2.0. In: Sixth ISCA Workshop on Speech Synthesis. pp. 294–299.

Zen, H., Tokuda, K., Black, A. W., 2009. Statistical parametric speech synthesis. Speech Commun. 51 (11), 1039–1064.