Speaking style conversion from normal to Lombard speech using a glottal vocoder and Bayesian GMMs

Ana Ramírez López, Shreyas Seshadri, Lauri Juvela, Okko Räsänen, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland.

ana.ramirez.lopez@aalto.fi, shreyas.seshadri@aalto.fi, lauri.juvela@aalto.fi, okko.rasanen@aalto.fi, paavo.alku@aalto.fi

Abstract

Speaking style conversion is the technology of converting natural speech signals from one style to another. In this study, we focus on normal-to-Lombard conversion. This can be used, for example, to enhance the intelligibility of speech in noisy environments. We propose a parametric approach that uses a vocoder to extract speech features. These features are mapped using Bayesian GMMs from utterances spoken in normal style to the corresponding features of Lombard speech. Finally, the mapped features are converted to a Lombard speech waveform with the vocoder. Two vocoders were compared in the proposed normal-to-Lombard conversion: a recently developed glottal vocoder that decomposes speech into glottal flow excitation and vocal tract, and the widely used STRAIGHT vocoder. The conversion quality was evaluated in two subjective listening tests measuring subjective similarity and naturalness. The similarity test results show that the system is able to convert normal speech into Lombard speech for the two vocoders. However, the subjective naturalness of the converted Lombard speech was clearly better using the glottal vocoder in comparison to STRAIGHT.
Index Terms: speaking style conversion, vocal effort, Lombard speech, glottal vocoder, Bayesian GMM

1. Introduction

Speaking style conversion is the technology of converting natural speech signals spoken in a particular style to another (e.g. whisper to shouting or normal to Lombard) while retaining the voice and linguistic information of the original speech signal. Speaking style conversion has multiple potential applications, such as personalizing speech to the needs of the end-listener and mapping speech that is difficult to understand in such a way that the signal becomes more intelligible. In the latter application, for example, normal speech could be converted into clear speech for hearing-impaired listeners. Similarly, people with normal hearing capacity could benefit from conversion of soft speech to a more intelligible style, such as Lombard speech [1], in noisy environments. It should be noted that in addition to keeping the linguistic and speaker information unchanged, a speaking style conversion system should not sacrifice speech quality. Therefore, this area of study calls for advanced technologies both in signal processing and machine learning. Speaking style conversion is related to other areas of speech technology such as statistical parametric speech synthesis (SPSS) [2], voice conversion (VC) [3], emotional voice conversion [4, 5] and speech intelligibility enhancement [6]. The topic can, however, be considered as a research area of its own because it differs from all the above areas: There is, for example, no linguistic-to-acoustic mapping as in speech synthesis and the conversion is not constrained by a strict latency requirement as in speech intelligibility enhancement. In the current study, we focus on converting normal speech to Lombard speech.

Compared to SPSS and VC, speaking style conversion has been studied only in a few previous investigations [7, 8, 9, 10], and the scope has been limited mainly to conversion of single words [8], isolated vowels [9], or logatomes (pseudo-words of one or many syllables) [7], rather than continuous speech. On the other hand, Lombard speech has been studied extensively in other areas of speech technology, such as SPSS [11] and intelligibility enhancement [12]. To our knowledge, the only previous study on normal-to-Lombard speaking style conversion was published in [8]. This study involves a rule-based solution that converts single words of normal speech to Lombard speech by modifying the original speech's fundamental frequency ($F0$), spectrum, and phoneme duration.

There are two main approaches to convert a source speech signal into a target one. One of them is a non-parametric approach that relies on processing directly the speech signal to achieve conversion, while the other one is a vocoder-based parametric approach, in which features are extracted with the vocoder, modified, and subsequently fed into the vocoder to synthesize the target speech signal. In this work, we choose to focus on a vocoder-based parametric approach where the vocoder is used to extract speech features both from the source and target styles and machine learning is used to learn a mapping between them.

The most widely used vocoder is STRAIGHT [13]. However, recent speech synthesis studies have shown that the so-called glottal vocoders constitute an effective alternative to STRAIGHT [14]. Given this, the goal of the current study is to build a speaking style conversion system and analyze its performance in normal-to-Lombard conversion by specifically exploring differences between STRAIGHT and a recent version of glottal vocoders [15]. Since the glottal vocoder aims to parameterize two main parts of natural speech production, the glottal excitation and vocal tract, we hypothesize the glottal vocoder to be a better vocoder candidate for the normal-to-Lombard conversion task. The speech features to be converted include the spectral tilt, $F0$, energy and duration, which are all known to be affected when natural talkers change their speaking style from normal to Lombard [16, 17]. To transform spectral and energy parameters, we employ Bayesian Gaussian mixture models (BGMMs) [18], while the duration mapping is achieved in a straightforward manner using frame-based interpolation of the vocoder features. BGMMs have the advantage of being less affected by overfitting than standard GMMs, which are used frequently in voice conversion [19]. This becomes particularly relevant in the current work, due to the limited availability of training data of Lombard speech. To the best of our knowledge, Bayesian extensions to standard GMMs have been applied previously in voice-conversion related research only in [20].

2. Speaking style conversion system

The speaking style conversion system is detailed in Figure 1. Prior to the actual conversion, the training is carried out as follows: First, a vocoder (STRAIGHT and the glottal vocoder studied here) is used to extract speech features (hereby denoted as vocoder features) at frame-level from both the source and target styles. Second, a mapping between the source and corresponding target features is learned for each of the selected vocoder features (here using BGMMs). Then, at the time of application: 1) Vocoder features are extracted from the given source-style speech signal, 2) the selected features are mapped to the target style, and 3) given all the vocoder features (the mapped features and the unmodified features), the vocoder synthesizes a speech signal in the required target style.

The current work aims to convert normal speech to Lombard speech by modifying the following attributes of the speech signal: 1) spectral tilt, 2) F0, 3) energy, and 4) duration of speech. Vocoder features representing the first three are mapped using one BGMM per feature. All three vocoder features were mapped for voiced frames, and energy was also mapped for (active) unvoiced frames. The voicing decision was made based on F0, while silent frames were detected using F0 and an energy threshold criterion. For training, the alignment of normal and Lombard frames was done using dynamic time warping (DTW) [21]. Aligned normal-Lombard frames that were in the opposite voiced/unvoiced categories were discarded from the training.

In order to modify the duration of the utterances, we scaled the duration of the voiced and unvoiced regions separately. The scaling was calculated as the mean ratio between the location of the aligned frames of the source and target styles in their corresponding utterance (outliers likely due to inaccuracies in DTW were removed). The scale values obtained were 1.08 and 0.88 for voiced and unvoiced regions respectively (in line with previous works [22]). The voiced and unvoiced regions were stretched and compressed, respectively, using frame-based interpolation. The duration modification was applied prior to BGMM mapping; non-converted features' duration was also modified. Furthermore, prior to synthesis, and to reduce distortions on the converted samples, the trajectories of the mapped features were smoothed with a moving average.

2.1. Vocoders

• ***Glottal vocoder -*** We use a recent variant of the glottal vocoder [14], which was originally developed for use in SPSS [2]. It uses quasi-closed phase (QCP) [23] glottal inverse filtering to decompose speech into a vocal tract filter and glottal flow excitation. Based on this, a deep neural network-based glottal pulse generation method was proposed in [15]. However, in contrast to text-to-speech, the present voice transformation task allows direct access to the original signal. Thus, in this work we use the vocoding procedure in [15] without any waveform modeling, but rather use the original estimated glottal waveforms as such. This is similar to linear prediction (LP) residual pitch-synchronous overlap-add (PSOLA) [24], and is not conventional vocoding in the sense that synthesis also uses non-parametric information. To parametrize speech, the following features are extracted with the vocoder: 1) log-energy, 2) harmonic-to-noise ratio (HNR), 3) F0, 4) vocal tract line spectral frequencies (LSFs), denoted here as LSF_{VT}, and 5) glottal source LSFs, denoted LSF_{glott}. The LSF_{glott} (for spectral tilt), $F0$ and energy vocoder features are chosen for the normal-to-Lombard conversion.

• ***STRAIGHT vocoder -*** This is a widely known speech vocoder that obtains a smooth spectral envelope such that the periodicity interference is minimized [13]. Here, the features extracted during analysis are: 1) the aperiodicity energy bands, 2) F0, and 3) the spectral envelope, represented through a Mel-generalized cepstrum (MGC). The features mapped are F0 and spectral tilt. Spectral tilt is modified by mapping the first two Mel cepstrum coefficients (c_1 and c_2) of the MGC feature, and keeping the other coefficients unchanged, as in [25]. Since the STRAIGHT vocoder does not include an explicit energy feature, energy adjustment was performed on the final synthesized speech signal. This was done at frame-level, and the signal was synthesized using overlap-add.

2.2. Bayesian GMM mapping

For the current work of style conversion, a BGMM between the two styles is trained for each vocoder feature. Vocoder features from the source style, $\mathbf{x}_s$, and target style, $\mathbf{x}_t$, are concatenated to obtain D-dimensional training data $\mathbf{x} = [\mathbf{x}_s,\ \mathbf{x}_t]^T$. Let $\mathbf{X} = [\mathbf{x}_1, .., \mathbf{x}_N]$ be modeled by a BGMM with K Gaussians with parameters $\{\theta_k\}_{i=1}^K$ and weights $\{\pi_k\}_{i=1}^K$, the likelihood of $\mathbf{X}$ is defined as

$$p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\theta}_k) \tag{1}$$

In the Bayesian setting we consider a prior on the model parameters and aim to infer their posterior distribution. The prior on the weights was chosen as the Dirichlet distribution i.e. $\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha}_0)$, where $\boldsymbol{\alpha}_0$ is a K-dimensional parameter. We consider full covariance Gaussians parameterized by the mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$, i.e. $\theta_k = \{\mu_k, \Lambda_k\}$. The conjugate prior is chosen for $\boldsymbol{\theta}$ as the Normal-Wishart distribution i.e. $\theta_k \sim \mathcal{NW}(\mathbf{m}_0, \beta_0, \mathbf{W}_0, \nu_0)$, where mean $\mathbf{m}_0$, scale matrix $\mathbf{W}_0$, real values $\beta_0 > 0$ and $\nu_0 > D - 1$ are parameters of the $\mathcal{NW}$ distribution [18]. Latent variables $\{z_i\}_{i=1}^N$ denote the Gaussian to which each of the N data points $\{\mathbf{x}_i\}_{i=1}^N$ are assigned.

There is no direct analytic solution for the posterior distribution of the BGMM parameters. This paper uses variational inference method [18] that approximates the analytically intractable posterior with a tractable distribution called variational distribution $q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$. This is done by making the following independence assumption:

$$q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \approx q(\mathbf{z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{z})q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \tag{2}$$

Kullback–Leibler (KL) divergence to the true posterior is then minimized to find the variational distribution. Since we use conjugate priors, $q(\boldsymbol{\pi})$ is another Dirchlet distribution $Dir(\boldsymbol{\alpha})$, and $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ another Normal-Wishart distribution $\mathcal{NW}(\mathbf{m}_k, \beta_k, \mathbf{W}_k, \nu_k)$ [18]. In practice, the final update equations are similar to the expectation–maximisation (EM) algorithm that iterates between finding the probabilities $q(\mathbf{z})$ (called responsibilities) based on the current model $q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, and updating model parameters based on the current responsibilities.

During application, the new source vocoder feature, $\mathbf{y}_s$, needs to be mapped to the target, $\mathbf{y}_t$. Let us first calculate the probability of data $\mathbf{y} = [\mathbf{y}_s,\ \mathbf{y}_t]^T$ given data $\mathbf{X}$ (modeled by the BGMM), $p(\mathbf{y}|\mathbf{X})$, called as the posterior predictive

$$p(\mathbf{y}|\mathbf{X}) = \frac{1}{\hat{\alpha}} \sum_{k=1}^{K} \alpha_k S_t(\mathbf{y}|\mathbf{m}_k, \boldsymbol{\Sigma}_k, \nu_k + 1 - D)$$
$$\text{where, } \boldsymbol{\Sigma}_k = \frac{1 + \beta_k}{(\nu_k + 1 - D)\beta_k} \mathbf{W}_k^{-1} \tag{3}$$

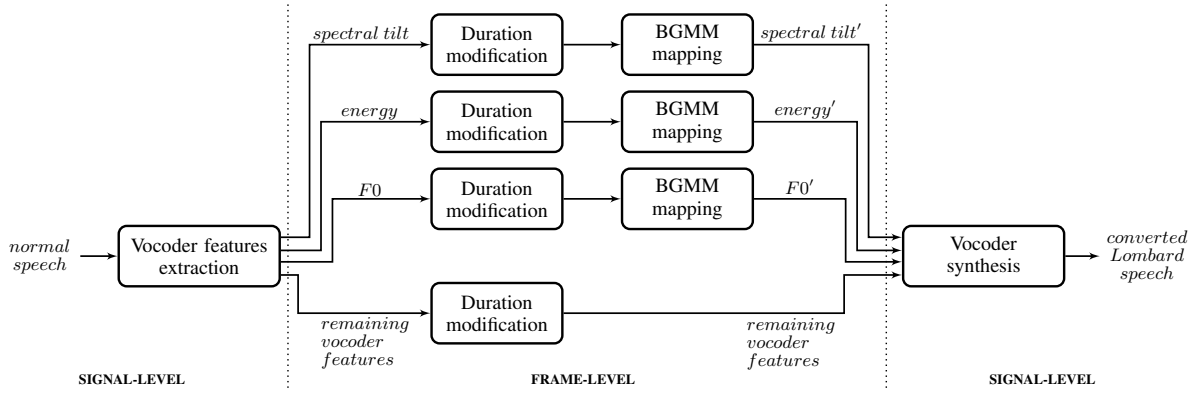

Figure 1: *Block diagram of the proposed speaking style conversion system. Prior to the conversion, the Bayesian Gaussian mixture models (GBMMs) are trained using pairs of normal and Lombard speech utterances.*

That is, a mixture of multivariate Student's t-distributions S_t with kth component having means $\mathbf{m}_k$ and covariance $\boldsymbol{\Sigma}_k$; and α_k is the kth term in $\boldsymbol{\alpha}$ and $\hat{\alpha} = \sum_k \alpha_k$ [18].

Let us consider the parameters of the kth multiavariate Student's t in Eq. (3) as block matrices $\mathbf{m}_k = [\mathbf{m}_s, \mathbf{m}_t]^T$ and $\boldsymbol{\Sigma}_k = \left[\begin{smallmatrix} \boldsymbol{\Sigma}_{ss} & \boldsymbol{\Sigma}_{st} \\ \boldsymbol{\Sigma}_{ts} & \boldsymbol{\Sigma}_{tt} \end{smallmatrix}\right]$. Now the MMSE estimate of $\boldsymbol{x}_t$ can be calculated, similar to a GMM mapping [26], as

$$\hat{\mathbf{y}}_t = \sum_{k=1}^{K} p(k|\mathbf{y}_s, \mathbf{X})[\mathbf{m}_t + \boldsymbol{\Sigma}_{ts}\boldsymbol{\Sigma}_{ss}^{-1}(\mathbf{y}_s - \mathbf{m}_s)] \quad (4)$$

where $p(k|\mathbf{y}_s, \mathbf{X})$ is the marginal probability of the kth component in Eq. (3), and the other term is the mean of the kth component in the conditional over the posterior predictive in Eq. (3) (see Section 10.7 of [27]). MATLAB codes for the BGMM mapping are available under an open source license[1].

3. Experimental setup

3.1. Data

Recordings from 10 Finnish speakers [28], with 4 female and 6 male speakers, were used for the current study. The recordings involved each speaker reading a text of 90 words, approximately one minute in duration. The same text was produced in two speaking styles, normal and Lombard speech. In order to elicit Lombard speech, the speakers heard background noise in their headphones while they were being recorded [28]. The recordings of each speaker, split into 11 utterances for each speaking style and down-sampled from 48 kHz to 16 kHz, were used in our experiments.

3.2. Normal-to-Lombard speech conversion

During feature extraction, analysis frames of 25 ms with a 5-ms frame shift were employed. For the glottal vocoder, the LSF_{glott} and LSF_{VT} features were 10 and 30-dimensional, respectively, the HNR feature consisted of 5 frequency channels, and F0 (as well as the glottal closure instants used in QCP) was computed using the REAPER tool [29]. In the STRAIGHT vocoder, the features consisted of 21 aperiodicity energy bands and the first 40 MGC coefficients (without the log-energy coefficient c_0). As in the glottal vocoder, F0 was extracted using REAPER [29]. The durations were modified using cubic spline interpolation for all features of the two vocoders, except for glottal vocoder's glottal excitation pulses, where nearest neighbour interpolation was applied.

[1] https://github.com/shreyas253/BGMM_Mapping

For the mapping, BGMMs were trained for each speaker and each vocoder feature using the utterances of the remaining speakers in the dataset (both females and males) as training data. Specifically, frame pairs of normal and Lombard speech of the corresponding feature to be mapped were used in training each BGMM. Since the Bayesian approach does not suffer from overfitting with even a large number of Gaussians, this number was fixed to K = 100 for all the vocoder features as significant improvements in terms of root-mean-square (RMS) error were not observed for larger values during a separate 10-fold cross-validation. Furthermore, the BGMM component means and precisions were modelled with prior distribution $\mathcal{NW}(\boldsymbol{\mu}_0, \beta_0, \mathbf{W}_0, \nu_0)$, whose parameters were set similar to those recommended in [30]: $\boldsymbol{\mu}_0$ and $\mathbf{W}_0$ were set to the dataset mean and precision, $\beta_0 = 1$, and $\nu_0 = D + 2$. The concentration parameter $\boldsymbol{\alpha}_0$ was set to the all ones vector.

3.3. Evaluation

Two listening tests were conducted to evaluate the quality of the samples obtained with the conversion system for the two vocoders, using the modified BeaqleJS evaluation framework [31]. 13 listeners took part on the first test, while 12 of the same listeners took part on the second test; all the listeners were Finnish natives.

The first evaluation was a similarity test, in which the perceptual similarity between the converted Lombard speech (vocoded either with the glottal vocoder or STRAIGHT) and natural Lombard speech was evaluated. The listeners were asked to rate, using a continuous scale from 1 to 5, how much a converted speech sample resembles a natural Lombard speech sample (1: none, 2: little, 3: moderately, 4: much, 5: very much). In rating the test sample, the listeners were given a non-converted reference which was generated by vocoding the corresponding sentence produced using normal speaking style. The listener was allowed to listen to the samples as many times as he/she wished. For this task, 16 utterances were randomly selected from the dataset (4 females and 4 males; 2 utterances per speaker). Therefore, since the listeners rated the conversion system for the two vocoders, each listener rated 32 test cases, which were presented in random order. Prior to the actual test, each listener had a training session in order to familiarize him/her with Lombard speech. In this training session, a subject was able to listen to a few sample pairs of normal vs. Lombard speech. The utterances of the training session were not used later in the test. Furthermore, the listeners were asked to adjust

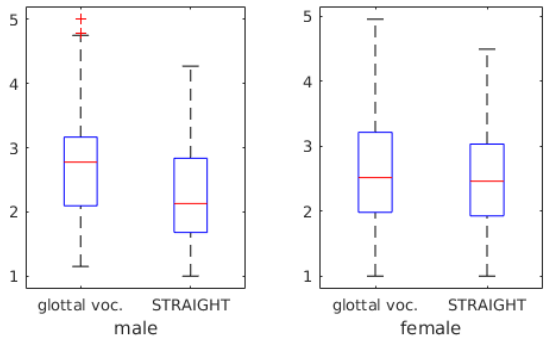

Figure 2: *Similarity test results, given in a scale from 1 to 5 that rates the resemblance of the converted sample to Lombard speech (1: none, 2: little, 3: moderately, 4: much, 5: very much).*

the volume to a loud yet comfortable level during the training session and to keep the chosen volume unchanged during the actual test.

The second evaluation was a pairwise comparison test, in which the naturalness of converted Lombard speech samples from the glottal vocoder and STRAIGHT were compared by the listeners. In this evaluation, the subjects listened to two versions of the same sentence, denoted as A and B, that represented the conversion conducted using the glottal vocoder or STRAIGHT, presented in a random order. The listener was asked which one sounds more natural. In addition, the listener was allowed to indicate if he/she had no preference to either. The listener could listen to the samples as many times as he/she wished. In this task, the listeners evaluated 24 test cases; the 24 utterances were selected randomly from the dataset (4 females and 4 males; 3 utterances per speaker).

4. Results

A boxplot of the similarity test results is shown in Figure 2: the central red line indicates the median, the boxes' edges are the 25th and 75th percentile, and the whiskers extend to the most extreme data points. Outliers are marked as red crosses. These results reveal that the speaking style conversion system is able to transform normal speech towards Lombard speech for the two vocoders. However, there are some gender specific distinctions: the median rate for the glottal vocoder-based system is slightly larger than the median rate for STRAIGHT in case of male speakers, while for female speakers the median rates have almost the same value.

Table 1 shows the results of the pairwise comparison test for naturalness of the speech; the results are shown as percentages of preference between the converted samples based on the glottal vocoder and STRAIGHT. The results show that the converted samples from the glottal vocoder case were clearly preferred in terms of naturalness (98.61% for males, and 97.92% for females) over those of STRAIGHT. Furthermore, there was also a very small number of cases in which the listeners had no preference between the two vocoders.

5. Conclusions

In this work, we proposed a speaking style conversion system to perform conversion from normal speech (source speaking style) to Lombard speech (target speaking style). In this system, a normal speech sample is converted by mapping (a selected set of) its speech features, extracted with a vocoder, into the corresponding features of Lombard speech using BGMMs, and subsequently using the vocoder with these features to synthesize speech in the target speaking style (Lombard speech). The conversion system involved a recently developed glottal vocoder that decomposes speech into a vocal tract filter and glottal flow excitation. This vocoder was compared in the proposed normal-to-Lombard speech conversion to the widely used STRAIGHT vocoder.

Table 1: *Results of preference task on naturalness, presented in percentages [%]. No pref. stands for 'No preference'.*

	Glottal vocoder	STRAIGHT	No pref.
Male	98.61	0.69	0.69
Female	97.92	0.00	2.08

Two subjective listening tests were employed to evaluate the conversion quality of the proposed system for the two vocoders. First, a similarity test evaluated the resemblance of the converted Lombard speech to natural Lombard speech. The results revealed that the conversion system was able to achieve conversion from normal to Lombard speech for the two vocoders. Both vocoders achieved the same level of resemblance to natural Lombard speech for female speakers. However, for males, the glottal vocoder was rated higher in resemblance than STRAIGHT. A possible explanation for this is that male voices are generally easier to parameterize accurately with glottal vocoders than female voices due to their lower pitch [32, 33], which makes the estimation of the glottal source with QCP more accurate. Second, a preference task compared the naturalness of the converted samples from both vocoders. The results showed that the converted samples obtained with the glottal vocoder were clearly more natural than those obtained with STRAIGHT.

While both vocoders managed to obtain similar ratings for Lombard-likeness of the speech, the converted samples from STRAIGHT vocoder presented artefacts (such as buzzing) that were more disruptive to the human ear in terms of naturalness than the artefacts present on the converted samples with the glottal vocoder. This would partly explain the clear preference of the listeners towards the glottal vocoder's samples.

Finally, it should be noted that the similarity test results showed that the rate of resemblance to natural Lombard speech of the converted Lombard samples did not reach a high level for any of the vocoders. The difference between natural normal speech and natural Lombard speech is prominent and there are many acoustical properties that change from one style to another. In the present work, the features selected for conversion were spectral tilt, F0, energy, and duration. The changes in the vocal tract are also key in Lombard speech, but these were not included in the current system to maintain simplicity. In consequence, further studies should involve vocal tract modifications that might increase the resemblance of the converted samples towards Lombard speech. In addition, while the BGMMs used in the present study provide a robust alternative for the mapping between speaking styles when the amount of training data is limited, the use of other alternative methods such as standard GMMs and DNNs should be explored and compared in the future together with larger amounts of training data.

6. Acknowledgements

This work was supported by the Academy of Finland (project numbers 274479 and 284671).

7. References

[1] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech, Language, and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.

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

[3] Y. Stylianou, "Voice transformation: a survey," in *Proc. of ICASSP*, Taipei, Taiwan, 2009, pp. 3585–3588.

[4] K. R. Scherer and T. Bänziger, "Emotional expression in prosody: A review and an agenda for future research," in *Proc. of Speech Prosody*, Nara, Japan, 2004, pp. 359–366.

[5] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[6] P. C. Loizou, *Speech enhancement: Theory and practice*. CRC press, 2013.

[7] À. Calzada and J. C. Socoró, "Vocal effort modification through harmonics plus noise model representation," in *Proc. of Nonlinear Speech Processing (NOLISP)*, Las Palmas de Gran Canaria, Spain, 2011, pp. 96–103.

[8] D.-Y. Huang, S. Rahardja, and E. P. Ong, "Lombard effect mimicking," in *Proc. of Speech Synthesis Workshop (SSW)*, Kyoto, Japan, 2010, pp. 258–263.

[9] K. I. Nordstrom, G. Tzanetakis, and P. F. Driessen, "Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1087–1096, 2008.

[10] C. d'Alessandro and B. Doval, "Experiments in voice quality modification of natural speech signals: the spectral approach," in *Proc. of ESCA/COCOSDA Workshop on Speech Synthesis*, Blue Mountains, Australia, 1998, pp. 277–282.

[11] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *Proc. of Blizzard Challenge 2010 Workshop*, Kansai Science City, Japan, 2010.

[12] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[13] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "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 Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[14] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, 2011.

[15] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *Proc. of ICASSP*, Shanghai, China, 2016, pp. 5120–5124.

[16] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[17] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, no. 1, pp. 151–173, 1996.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[19] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[20] L. Li, Y. Nankaku, and K. Tokuda, "A bayesian approach to voice conversion based on gmms using multiple model structures," in *Proc. of Interspeech*, Florence, Italy, 2011, pp. 661–664.

[21] D. Ellis, "Dynamic time warp (DTW) in Matlab," http://www.ee.columbia.edu/dpwe/resources/matlab/dtw/, 2003.

[22] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 366–378, 2009.

[23] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 596–607, 2014.

[24] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175–205, 1995.

[25] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Mel cepstral coefficient modification based on the glimpse proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise," in *Proc. of Interspeech*, Portland, USA, 2012.

[26] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. of ICASSP*, Seattle, USA, 1998, pp. 285–288.

[27] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," Tech. Rep., 2007.

[28] E. Jokinen, U. Remes, and P. Alku, "The use of read versus conversational Lombard speech in spectral tilt modeling for intelligibility enhancement in near-end noise conditions," in *Proc. of Interspeech*, San Francisco, USA, 2016, pp. 2771–2775.

[29] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," https://github.com/google/REAPER, 2015.

[30] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[31] S. Kraft and U. Zölzer, "BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in *Proc. of Linux Audio Conference*, Karlsruhe, Germany, 2014.

[32] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM entry for Blizzard Challenge 2011: Utilizing source unit selection in hmm-based speech synthesis for improved excitation generation," in *Proc. of Blizzard Challenge 2011 Workshop*, Turin, Italy, 2011.

[33] ——, "The GlottHMM entry for Blizzard Challenge 2012: Hybrid approach," in *Proc. of Blizzard Challenge 2012 Workshop*, Portland, USA, 2012.