Seshadri, Shreyas; Juvela, Lauri; Räsänen, Okko; Alku, Paavo

Vocal Effort Based Speaking Style Conversion Using Vocoder Features and Parallel Learning

Published in:
IEEE Access

DOI:
10.1109/ACCESS.2019.2895923

Published: 01/01/2019

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 9, 2019, accepted January 23, 2019, date of publication January 31, 2019, date of current version February 14, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2895923

Vocal Effort Based Speaking Style Conversion Using Vocoder Features and Parallel Learning

SHREYAS SESHADRI[1], LAURI JUVELA[1], OKKO RÄSÄNEN[1,2], AND PAAVO ALKU[1], (Senior Member, IEEE)

[1]Department of Signal Processing and Acoustics, Aalto University, FI-00076 Espoo, Finland
[2]Laboratory of Signal Processing, Tampere University of Technology, 33720 Tampere, Finland

Corresponding author: Shreyas Seshadri (shreyas.seshadri@aalto.fi)

This work was supported by the Academy of Finland, under Grant 312105, Grant 312490, and Grant 314602.

ABSTRACT Speaking style conversion (SSC) is the technology of converting natural speech signals from one style to another. In this study, we aim to provide a general SSC system for converting styles with varying vocal effort and focus on normal-to-Lombard conversion as a case study of this problem. We propose a parametric approach that uses a vocoder to extract speech features. These features are mapped using parallel machine learning models 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. A total of three vocoders (GlottDNN, STRAIGHT, and Pulse model in log domain (PML)) and three machine learning mapping methods (standard GMM, Bayesian GMM, and feed-forward DNN) were compared in the proposed normal-to-Lombard style conversion system. The conversion was evaluated using two subjective listening tests measuring perceived Lombardness and quality of the converted speech signals, and by using an instrumental measure called Speech Intelligibility in Bits (SIIB) for speech intelligibility evaluation under various noise levels. The results of the subjective tests show that the system is able to convert normal speech into Lombard speech and that there is a trade-off between quality and Lombardness of the mapped utterances. The GlottDNN and PML stand out as the best vocoders in terms of quality and Lombardness, respectively, whereas the DNN is the best mapping method in terms of Lombardness. PML with the standard GMM seems to give a good compromise between the two attributes. The SIIB experiments indicate that intelligibility of converted speech compared to that of normal speech improved in noisy conditions most effectively when DNN mapping was used with STRAIGHT and PML.

INDEX TERMS Bayesian GMM, DNN, GlottDNN, Lombard speech, pulse model in log domain, speaking style conversion, vocal effort.

I. INTRODUCTION

Speaking style conversion (SSC) is the technology of converting natural speech signals spoken in a particular style to another (e.g. whisper-to-normal or normal-to-Lombard [1]) while retaining the linguistic and speaker-specific information of the original speech signal. SSC 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 [2], [3] 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, in noisy environments. It should be noted that in addition to keeping the linguistic and speaker-specific information unchanged, an SSC system should not sacrifice speech quality. Therefore, this area of study calls for advanced technologies both in signal processing and machine learning.

SSC is related to other areas of speech technology such as statistical parametric speech synthesis (SPSS) [4], voice conversion (VC) [5], and speech intelligibility enhancement in speech transmission [6]. The topic can, however, be considered as a research area of its own because it differs from all the above areas. For example, there is no linguistic-to-acoustic mapping as in SPSS. Furthermore, the strict latency requirements of speech intelligibility enhancement in speech

The associate editor coordinating the review of this manuscript and approving it for publication was Berdakh Abibullaev.

transmission [7] are not necessarily present in SSC, where offline processing is also possible for several potential use scenarios. At the broadest level, SSC involves conversions of any properties of speech that do not carry immediate lexical, syntactic, or speaker identity related information without sacrificing the original quality of the signal.

An example of SSC is emotion conversion (e.g., [8]–[15]), which is a topic in which the speaking style conversion is conducted on affective, paralinguistic properties of speech. Another key dimension of speaking style is related to the vocal effort used by the speaker in different social and communication contexts to meet the desired communicative goals. These speaking style modifications are typically associated with changes in signal intensity, loudness (that takes into account spectral balance), and pitch with the aim of maintaining speech intelligibility in different noise conditions or at different spatial distances between the speaker and the listener. Different levels of vocal effort can be seen as forming a continuum consisting of *whispered*, *normal*, *loud*, *Lombard*, and *shouted* speech (see also work on *clear* speech; [16], [17]), even though there are several subtle articulatory and acoustic differences between these styles that do not follow simple linear relationships. Since speech recording and reproduction environments (or the original and new target listeners) may differ from each other, and since the different styles on the above-mentioned continuum are directly related to the spoken communication's success and suitability, it would be beneficial to be able to tailor the speech signal along this continuum through the use of SSC technology. While there has already been work in whispered-to-normal speech conversion (e.g., [18]–[22]), SSC for other aspects of vocal effort has only been studied in a small number of previous works [23]–[27].

In the current study, we focus on converting normal speech to Lombard speech. Lombard speech corresponds to a speaking style that talkers naturally employ in noisy environments to improve intelligibility, and it has been studied extensively in other areas of speech technology such as SPSS [28], speaker recognition [29] and intelligibility enhancement [30]. To our knowledge, the only previous studies on normal-to-Lombard SSC were [23], which involved the use of non-uniform time scale modification, formant shifting, and energy redistribution in the presence of car noise, and [25] that involved a rule-based solution that converts single words from normal to Lombard speech by modifying the natural speech signal's fundamental frequency ($F0$), spectrum, and phoneme duration.

Modern speech technology systems are typically data driven, where speech data needs to be collected to train a machine learning mapping between source and target representations. However, collection of a large quantity of Lombard speech data (as well as data from some other styles along the vocal effort continuum such as shouted speech) is laborious and potentially injurious to health of the speakers. This data sparsity limits the straightforward use of approaches similar to the recent end-to-end TTS systems, that are able to learn global style tokens for controlling the speaking style [31]. These types of systems typically consist of two highly data-hungry components: a sequence-to-sequence mel-spectrogram predictor and a WaveNet vocoder [32]. While the Tacotron line of work [31], [32] has not reported their training dataset size, for example [33] used 20 hours of data for training a speaker dependent model. For training WaveNets, [34] recommended at least 3 hours of data for a single speaker TTS system, while [35] trained a multi-speaker WaveNet vocoder with approximately one hour of data per speaker. In the present study, we investigate a scenario where we have access to only a few *minutes* of recorded speech data in both normal and Lombard styles (see Section VI-A). To handle the data sparsity, the present normal-to-Lombard parametric SSC system relies on *parametric vocoders* (VOCs) for feature extraction and *machine learning models* (MLMs) for speech modification. The system utilizes parallel training where the MLMs are trained with speech utterance pairs from source and target styles having the same linguistic content and speaker.

The normal-to-Lombard SSC system of this study is based on modifying the most important features between the two styles: the fundamental frequency (F0), spectral tilt, signal energy, and segment duration. For this purpose, we investigate the use of three vocoders familiar from SPSS: GlottDNN [36], STRAIGHT [37], and Pulse model in log domain (PML) [38] (see Section III for details). Related to MLMs, we compare standard Expectation-maximization (EM) [39], [40] algorithm-based Gaussian mixture models (SGMMs), Bayesian Gaussian mixture models (BGMMs), and feed-forward deep neural nets (DNNs), as they provide a range of approaches from the standard methods to more recent and popular methods used in similar problems (see Section IV for details). Hence, we explore in total 9 different combinations of VOCs and MLMs in the present study.

The overall goal of this study is to first provide the general framework for an SSC system, and then implement the particular case of a normal-Lombard SSC using VOCs, MLMs, and parallel training. The system is trained on Finnish recordings from [41] and evaluated using subjective listening tests, instrumental speech intelligibility experiments in noisy conditions as well as with objective analysis of the distributions of the mapped features.

The paper is organized as follows: Section II describes the general structure and outline of an SSC system. Sections III and IV provide a basic framework for the VOCs and MLMs used in the current study. Then Section V details our specific case of a normal-to-Lombard SSC system. Section VI explains the experimental setup, including data used, model adaptation techniques, and system specification details. And finally Sections VII, VIII, and IX describe the evaluation, results, and discussion and conclusions, respectively.

II. PARAMETRIC SPEAKING STYLE CONVERSION

In principle, the SSC problem could be approached as a direct transformation (such as filtering) or as an end-to-end

mapping problem where the original speech waveform or its full-band spectral representation is directly transformed into the target style. An example of the direct transformation approach is [26], which uses adaptive pre-emphasis linear prediction to transform the signal in terms of its vocal effort and breathiness.

However, an alternative approach for SSC is to use a *parametric* technique, hereby referred to as *parametric speaking style conversion*, in which selected speech features are first extracted, modified, and finally used to synthesize the speech signal in the target style. The potential advantage of the parametric approach over direct processing is that the former enables the combination of machine learning methods with a priori knowledge of the speech production system, enabling parametrization and modification of key properties of the signal that are related to the phenomenon of interest while keeping other aspects of the signal intact. This allows the training of SSC systems for conversion problems where a limited amount of training data are available, as the key parameters can be modeled (partially) independently of the factors that are not related to the conversion task. The parametric approach also provides better manual control and interpretability of the conversion system behavior. The obvious drawbacks are the potentially erroneous assumptions about the independence of features w.r.t. the style dimension of interest and erroneous extraction of speech features due to, for example, high pitch or lack of voicing. In addition, the parametric approach might suffer from challenges in fusing the different features back into high-fidelity waveforms when only some of them have been transformed, and from the need for a priori knowledge for selecting the features of interest for the mapping. This calls for efficient methods for parametrizing and synthesizing the acoustic signal and for robust methods for learning the feature mappings between the styles of interest.

The general structure of a parametric SSC system is shown in Figure 1. The input to the system is a speech utterance spoken in the source style, and the output is the same utterance in the desired target style. The system consists of three main parts: *feature extraction*, *mapping model*, and *synthesis*. In the first part, all the features that are necessary for the speech synthesis part of the system are first extracted from the input signal. The features that are known to contribute most to the source-to-target style conversion in question are then converted in the second part of the system using a mapping model. This mapping model can in principle be either supervised (i.e., previously trained on data from both of the styles in question) or unsupervised. Finally, in the third part of the system, the mapped as well as unmodified features are fed to the synthesis system, which generates the speech signal in the desired target style.

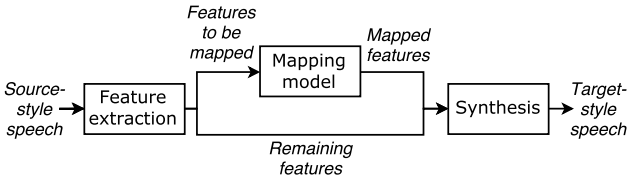

FIGURE 1. Block diagram of the parametric speaking style conversion system.

By referring to Figure 1, the SSC system implemented in the current work utilizes VOCs for feature extraction and synthesis, and MLMs in a supervised manner for the mapping models. The technologies used in the current study to implement VOCs and MLMs are described next in more detail in sections III and IV, respectively.

III. VOCODERS

Vocoders are widely used particularly in SPSS to express speech in parametric forms. Based on the parameterization scheme used, vocoders can be categorized to glottal vocoders (e.g. GlottHMM [42], GlottDNN [36]), mixed/impulse excited vocoders (e.g. STRAIGHT [37], WORLD [43]), sinusoidal vocoders (e.g. Quasiharmonic model [44] , dynamic sinusoidal model [45]) and source-filter vocoders that use sinusoidal signal analysis as a measure of harmonicity (PML [38]). The glottal and mixed/impulse excited vocoders utilize the source-filter model of speech production [46], which assumes that speech is produced by a source signal that is convolved with a filter conveying the vocal tract formants. The mixed excitation approach assumes that the excitation is spectrally flat and contains the pitch, noise, and phase information, and the filter models the entire spectral envelope of the signal. In glottal vocoders, the excitation of voiced speech is a model of the true acoustical source generated by the vocal folds, the glottal volume velocity waveform. The spectral envelope of this excitation is not flat but shows a tilt that varies, for example, based on the vocal effort or phonation type. Finally, sinusoidal vocoders represent speech as a sum of sinusoidal functions, evolving over time. In order to analyze the potential differences between vocoders for the current task of SSC, we analyze three vocoders from different categories: GlottDNN, which was shown in a recent study [36] to be the most potential glottal vocoder, STRAIGHT [37], which is the most widely used vocoder in SPSS, and PML [38], which demonstrated good performance in two recent vocoder studies [36], [38]. Sections III-A, III-B and III-C, provide a brief look and the structures of each of these vocoders.

A. GLOTTDNN

The GlottDNN [36] (based on the earlier implementation GlottHMM [42]), uses a quasi-closed phase (QCP) [47] glottal inverse filtering to decompose speech into a vocal tract filter and glottal flow excitation. During synthesis, GlottDNN uses a feed-forward DNN trained on acoustic features to generate the glottal pulses (as proposed in [48]). GlottDNN models voiced segments of speech as a convolution of the glottal flow excitation and vocal tract which are estimated using the QCP glottal inverse filtering algorithm [47]. Unvoiced segments are modeled with random noise excitation and

conventional linear prediction (LP) model for the vocal tract. To parametrize speech, the following features are extracted in the current study when using GlottDNN as VOC: 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 envelope LSFs, denoted here as LSF_{glott}.

B. STRAIGHT

STRAIGHT [37], [49] is a widely used source-filter vocoder that models speech spectrum as a smooth envelope (filter) that is excited with a spectrally flat excitation signal (source). The spectral envelope is estimated using two pitch-adaptive analysis windows (primary and complementary). The primary window consists of a Gaussian window convolved with a triangular B-spline window, while an asymmetric complementary window is created by multiplying the primary window with a sine window [37]. Finally, the spectral estimates obtained with these windows are combined by taking a weighted quadratic mean, such that the chosen weight minimizes the harmonic interference in the resulting time-frequency envelope [37]. The aperiodicity spectrum is estimated by comparing upper and lower spectral envelopes [49]. In synthesis stage, aperiodicity is used to modify a periodic impulse train excitation to create a mixed excitation signal for voiced speech, while white Gaussian noise excitation is used for unvoiced speech. For parametric processing, we represent the envelope with mel-generalized cepstral coefficients (MGCs, [50]) and the aperiodicity spectrum by log-averages over equivalent rectangular bandwidth (ERB) auditory bands. Thus, the features extracted for parameterization are: 1) the aperiodicity band energies (BAP), 2) F0, and 3) the spectral envelope MGCs.

C. PULSE MODEL IN LOG DOMAIN

The Pulse model in log-domain (PML) [38] is a recent state-of-the art vocoder utilizing sinusoidal signal analysis and pitch synchronous pulse-based synthesis. PML supports using generic spectral envelopes, and Degottex *et al.* [38] recommend using the STRAIGHT or WORLD envelope. The vocoder's distinctive property is its aperiodicity modeling via a phase distortion deviation (PDD) spectrum, which generalizes to modeling both voiced and unvoiced speech without explicit voicing decisions. The PDD is thresholded to produce a binary noise mask (BNM), which is averaged in mel-bands for parametric processing. Here, the features extracted during analysis are: 1) the binary noise mask (BNM), 2) F0, and 3) the spectral envelope.

IV. MACHINE LEARNING MAPPING METHODS

As for MLMs, three different techniques are explored in the current study: an SGMM, a BGMM, and a feed-forward DNN. SGMMs are used frequently in related fields such as VC (e.g. [51]) and intelligibility enhancement (e.g. [52], [53]). The BGMMs are a Bayesian extension to the SGMMs, which could be potentially beneficial in scenarios with limited data due to their capability to scale model complexity to the structure of the data. To the best of our knowledge, Bayesian extensions to standard GMMs have been applied previously in voice-conversion related research only in [54]. DNNs, on the other hand, are some of the most widely used MLMs in recent years across a large number of domains. DNNs, although extremely powerful, may also suffer more from limited training data in comparison to the classical mixture models that typically have much fewer parameters to estimate. Hence, the current problem of SSC offers a good opportunity to compare these methods.

The technical difference between SGMMs and BGMMs is discussed in detail in the following Sections IV-A and IV-B, while the practical implementation details of those and the DNN can be found in Section VI-C4. The source codes of the three MLMs used in this paper are available under an open source license[1] for reference and reproducibility.

A. STANDARD GAUSSIAN MIXTURE MODELS (SGMM)

A mapping model is trained on the data set consisting of vocoder features from both the source and target style. Let us consider the training set consisting of N vocoder feature vectors of dimensionality $D/2$ from the source style $\mathbf{x}_s$ and target style $\mathbf{x}_t$. During application, the new source data, $\mathbf{y}_s$, needs to be mapped to the target, $\mathbf{y}_t$, using the trained mapping model.

In the standard GMM approach, the training set of the source, $\mathbf{x}_s$, and target data, $\mathbf{x}_t$, is concatenated as $\mathbf{x} = [\mathbf{x}_s^T, \mathbf{x}_t^T]^T$ to obtain N samples of D-dimensional training data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ for the SGMM model. Let $\mathbf{X}$ be modeled by an SGMM with K full covariance Gaussians, each having parameters $\{\boldsymbol{\theta}_k\}_{i=1}^K$ and weights $\{\pi_k\}_{i=1}^K$. In the current, frequentist interpretation of the GMM, the parameters are considered as fixed values to be estimated by maximizing the likelihood of $\mathbf{X}$ defined as

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

where the parameters, $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \Sigma_k\}$, are the mean and covariance of the kth Gaussian and the weights, π_k, sum to one. The values of the parameters that maximize the likelihood are found using the EM [39], [40] algorithm.

Let us consider the parameters of the kth Gaussian as block matrices corresponding to the source, s, and target, t, features as $\boldsymbol{\mu}_k = [\boldsymbol{\mu}_{s|k}, \boldsymbol{\mu}_{t|k}]^T$ and $\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_{ss|k} & \boldsymbol{\Sigma}_{st|k} \\ \boldsymbol{\Sigma}_{ts|k} & \boldsymbol{\Sigma}_{tt|k} \end{bmatrix}$. Now, during application, the minimum mean square error (MMSE) estimate of target features $\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t$ can be calculated as

$$\hat{\mathbf{y}}_t = \sum_{k=1}^{K} p(k|\mathbf{y}_s, \mathbf{X})[\boldsymbol{\mu}_{t|k} + \boldsymbol{\Sigma}_{ts|k}\boldsymbol{\Sigma}_{ss|k}^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_{s|k})] \tag{2}$$

where $p(k|\mathbf{y}_s, \mathbf{X})$ is the probability of the kth component calculated based on π_k and marginal likelihood of the kth Gaussian; and the other term is the mean of conditional likelihood of the kth Gaussian. (See [55] for a detailed

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

derivation and [52], [53], [55] for other use-cases of SGMM mapping.)

In case of additional training data for both source style $\mathbf{a}_s$ and target style $\mathbf{a}_t$ after we have trained the original SGMM model, model adaptation can be applied to update the model. This is often used in speaker verification [56]. Again the new training data is concatenated as $\mathbf{a} = [\mathbf{a}_s,\ \mathbf{a}_t]^T$ to get N' new training data samples $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_{N'}]$. We first calculate a set of sufficient statistics for each Gaussian component, k, of the original SGMM

$$\begin{aligned} n'_k &= \sum_{i=1}^{N'} P(k|\mathbf{a}_i) \\ E_k(\mathbf{a}) &= \frac{1}{n'_k} \sum_{i=1}^{N'} P(k|\mathbf{a}_i)\mathbf{a}_i \\ E_k(\mathbf{a}^2) &= \frac{1}{n'_k} \sum_{i=1}^{N'} P(k|\mathbf{a}_i)\mathbf{a}_i\mathbf{a}_i^T \end{aligned} \tag{3}$$

The parameters for each component, k, of the adapted SGMM are now calculated as

$$\begin{aligned} \hat{\pi}_k &= \delta_k \frac{n'_k}{K} + (1-\delta_k)\pi_k \\ \hat{\boldsymbol{\mu}}_k &= \delta_k E_k(\mathbf{a}) + (1-\delta_k)\boldsymbol{\mu}_k \\ \hat{\boldsymbol{\Sigma}}_k &= \delta_k E_k(\mathbf{a}^2) + (1-\delta_k)(\boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^T + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T) - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T \end{aligned} \tag{4}$$

where δ_k is the kth weighting factor influencing the effect of the old model. The weighting factor is chosen as $\delta_k = n'_k/(n'_k + n_k)$, where $n_k = \sum_{i=1}^{N} P(k|\mathbf{x}_i)$, calculated from the original SGMM model.

B. BAYESIAN GAUSSIAN MIXTURE MODELS (BGMM)

Similar to the SGMM, we concatenate $\mathbf{x} = [\mathbf{x}_s^T,\ \mathbf{x}_t^T]^T$ and let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ 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

$$p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\theta}_k) \tag{5}$$

In the Bayesian setting, we consider the model parameters as random variables with a prior distribution. We aim to infer their posterior distribution once we have observed the data. 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 [40]. 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 the variational inference method [40] that approximates the analytically intractable posterior with a tractable distribution called the variational distribution $q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$. This is done by making the following independence assumption:

$$q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \approx q(\mathbf{z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{z})q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \tag{6}$$

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 Dirichlet distribution $Dir(\boldsymbol{\alpha})$, and $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ is another Normal-Wishart distribution $\mathcal{NW}(\mathbf{m}_k, \beta_k, \mathbf{W}_k, \nu_k)$ (see [40] for details). In practice, the final update equations are similar to the EM algorithm [39], [40] used for the SGMM 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, in order to make predictions of the target features $\mathbf{y}_s$ from the source features $\mathbf{y}_t$, we need to consider the posterior predictive distribution of the BGMM. The posterior predictive distribution, $p(\mathbf{y}|\mathbf{X})$, of the data $\mathbf{y} = [\mathbf{y}_s,\ \mathbf{y}_t]^T$ given data $\mathbf{X}$ is given by

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) &= \frac{1}{\hat{\alpha}} \sum_{k=1}^{K} \alpha_k S_t(\mathbf{y}|\mathbf{m}_k, \boldsymbol{\Sigma}_k, \nu_k + 1 - D) \\ \text{where,} \quad \boldsymbol{\Sigma}_k &= \frac{1+\beta_k}{(\nu_k + 1 - D)\beta_k} \mathbf{W}_k^{-1} \end{aligned} \tag{7}$$

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 $\hat{\alpha} = \sum_k \alpha_k$ [40].

Now let us consider the parameters of the kth multivariate Student's t in Eq. (7) as block matrices $\mathbf{m}_k = [\mathbf{m}_s,\ \mathbf{m}_t]^T$ and $\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_{ss} & \boldsymbol{\Sigma}_{st} \\ \boldsymbol{\Sigma}_{ts} & \boldsymbol{\Sigma}_{tt} \end{bmatrix}$. Now the MMSE estimate of $\boldsymbol{y}_t$, $\hat{\mathbf{y}}_t$, can be calculated as in the SGMM

$$\hat{\mathbf{y}}_t = \sum_{k=1}^{K} p(k|\mathbf{y}_s, \mathbf{X})[\mathbf{m}_t + \boldsymbol{\Sigma}_{ts}\boldsymbol{\Sigma}_{ss}^{-1}(\mathbf{y}_s - \mathbf{m}_s)] \tag{8}$$

where $p(k|\mathbf{y}_s, \mathbf{X})$ is the marginal probability of the kth component in Eq. (7), and the other term is the mean of the kth component in the conditional over the posterior predictive in Eq. (7) (see [57, Sec. 10.7]). MATLAB codes for the BGMM mapping are available under an open source license.[2]

C. FEED-FORWARD DNN

As the third MLM alternative, standard feed-forward multilayer perceptron (MLP) DNNs [58] were used to train the mapping model with $\mathbf{x}_s$ as the inputs and $\mathbf{x}_t$ as the target. Technical details of the implementation are described in Section VI-C4, and an interested reader is referred to [58] for a basic description of MLPs.

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

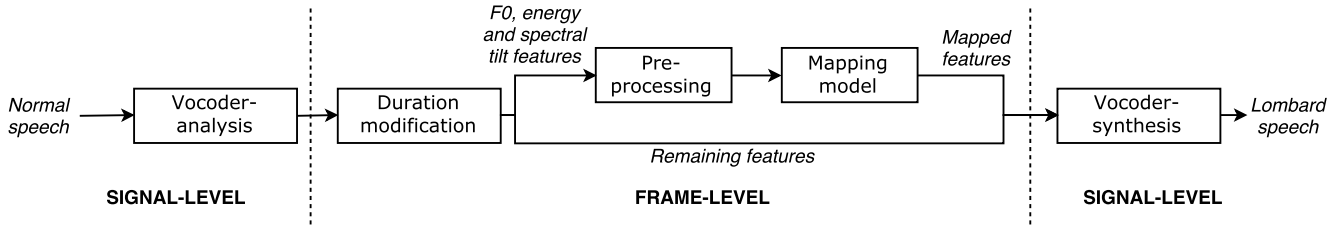

FIGURE 2. Block diagram of the proposed normal-to-Lombard SSC system. Prior to the conversion, the mapping models are trained using DTW aligned pairs of normal and Lombard speech utterances.

V. NORMAL-TO-LOMBARD SSC SYSTEM

Our normal-to-Lombard parametric SSC system is detailed in Figure 2. Prior to the actual conversion, the training of the MLMs is carried out. Firstly, a VOC is used to extract select speech features that are known to contribute to the specific style conversion in question (denoted as VOC features) at frame-level from both the source and target styles. Then, an MLM is trained between the source and corresponding target features for the pre-processed set of the selected vocoder features (see Section VI-C4 for details). Parallel training is employed here, i.e., a pair of source and target style utterances is used that have the same linguistic content and speaker. During the actual conversion: 1) all the VOC features necessary for synthesis are extracted from the given source-style speech signal, 2) duration modification is done for all features, 3) the selected features are mapped to the target style using the trained MLM, and 4) with all the modified VOC features, VOC synthesizes a speech signal in the required target style.

Several variations of features were tried for mapping. It was found that mapping all spectral features resulted in very poor perceptual quality of the generated speech signals for all the three vocoders considered in the present study. This is likely due to the very limited size of the training data that we have access to. Hence, the mapping was restricted to modifications of the features that are known to be most important to the Lombard style. The following attributes of the speech signal are modified to achieve the conversion from normal to Lombard style: 1) spectral tilt, 2) F0, 3) energy, and 4) duration. Vocoder features representing the first three are mapped using a trained model for voiced frames. For parallel training, the alignment of normal and Lombard speech frames is first performed using dynamic time warping (DTW) [59]. The voicing decision is made based on F0, while silent frames are detected using F0 and an energy threshold criterion.

Duration conversion of the utterances is done by scaling the duration of the voiced and unvoiced regions by a constant factor that is separately estimated for voiced and unvoiced segments. This is a simplistic approach and does not perfectly encapsulate the real world mechanism where the linguistic information influences the duration of segments (see [29] for an analysis of the duration of different phoneme-classes in Lombard speech). Ideally the duration modification of speech segments would also be modeled by an MLM that takes into account the linguistic identity and context of these units. However, training such a duration model would also require a notable increase in the amount of training data and diversity. Therefore, a constant-factor duration conversion is a good approximation applicable to a limited-data scenario such as the present study.

It should also be noted that, in addition to the four attributes described above, vocal tract modifications may also contribute to Lombard speech through a shifting of the formant frequencies and a narrowing of their bandwidths. Initial experiments were conducted by trying to directly map all vocoder features required for synthesis with the same procedure, but the perceptual quality of the resulting mapped utterances was either poor or the Lombard effect was much weaker than in case of the final feature set, depending on the model configuration. This is likely due to the limited data currently available for training the MLMs. The vocal tract information is inherently dependent on the speaker and the linguistic content of speech, requiring more data to learn a high-quality mapping without audible artifacts. Therefore, vocal tract mapping is not included in the present system, and new, more advanced parameterization methods of the vocal tract needs to be considered for SSC in the future.

VI. EXPERIMENTAL SETUP

A. DATA

Speech recordings from 10 Finnish speakers, 4 female and 6 male, were used to train the system and to carry out evaluations. The recordings (see [41] for details) 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. In order to elicit Lombard speech, background noise (highly unstationary pub noise [60], with A-weighted sound pressure level (SPL) of approximately 80 dB) was played to the speakers' ears with headphones while they were being recorded [41]. The recordings of each speaker were split into 11 lexically unique utterances (same utterances for each talker), which will from now on be referred to as sentences. The duration of the sentences in the two speaking styles varied between 2 and 9 seconds. Hence, our dataset consisted of 10 speakers, each speaking 11 sentences in both speaking styles, corresponding to a total of 220 utterances. These data were down-sampled from 48 kHz to 16 kHz to be used in the experiments of the current study.

B. MODEL ADAPTATION

In the experiments, a separate MLM was trained for each test speaker by using the utterances of the 9 other speakers in the dataset (both females and males). Since there was large variance in the degree of Lombardness in the speech of the talkers in our corpus, training equally on data from all talkers would have led to a perceptually mediocre Lombard effect in the synthesized speech. To avoid this, model adaptation was applied to the MLMs by first training them on the full set of talkers, and then further adapting these models to a subset of the original data from two handpicked talkers of the same gender with a larger and more pronounced Lombard effect (never using the same talker for adaptation and testing). This approach enabled us to train the models on the best subset of the training data (in terms of Lombardness) without risking substantial overfitting of the models.

Model adaptation for the SGMM described in Section IV-A is based on techniques used in speaker verification [56]. Adaptation is much simpler in the Bayesian setting. The posterior distribution of X calculated earlier for the weight distribution $q(\boldsymbol{\pi})$ and for the kth Gaussian $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ are now used as the priors for the new training data, A. The adapted BGMM posteriors are calculated as usual (see Section IV-B). During the model adaptation for DNNs, a new feed-forward DNN was trained with the same structure and specifications as the original having $\mathbf{a}_s$ as the inputs and $\mathbf{a}_t$ as the target. However, the weights were initialized to those optimized by the original DNN (see Section VI-C4 for details).

C. SYSTEM SPECIFICATIONS

This section details the hyper-parameters and specifications used for each step of the normal-to-Lombard SSC system shown in Figure 2. These values were set based on brief evaluations of the RMS errors and perceived quality of the final synthesized speech signals

1) VOCODERS

During feature extraction, analysis frames of 25 ms with a 5-ms frame shift were employed. F0 was computed using the RAPT algorithm from the SPTK toolkit [61], with the range of allowed frequencies set to 50–500 Hz. For GlottDNN, the LSF_{glott} and LSF_{VT} features were 10- and 30-dimensional, respectively. The HNR feature consisted of 5 frequency channels. The glottal closure instants used in QCP were computed using the REAPER tool [62] with the same allowed frequency range as F0. A 3-hidden layer feed-forward DNN with sigmoid activations and layer sizes of 150, 250, and 300 was trained for each speaker to generate the 400-dimensional glottal pulse waveform. The training was done with the GlottDNN features from the remaining 9 speakers. The DNNs were optimized based on mean squared error (MSE) using stochastic gradient descent with a mini-batch size of 100, learning rate of 0.01 and early stopping criterion with patience of 10 epochs, and maximum number of epochs set to 50. As mentioned in Section V, the features mapped are the F0, energy, and spectral tilt. These features correspond to frame-wise F0, energy, and LSF_{glott} features, respectively, extracted for GlottDNN. Ideally, the spectrum of the VT filter, represented by LSF_{VT}, shows local peaks, formants, but the overall spectral envelope of the filter is flat and the spectral tilt of the speech signal is modeled by the glottal filter, represented by LSF_{glott}. However, QCP analysis has a tendency to include some tilt in its VT estimate. This effect varies between speakers and may cause inconsistency issues when the MLMs are trained in cross-speaker fashion. To compensate, we parameterized the spectral tilt of the VT filter with a first order LP filter, removed the estimated tilt from the VT filter, and transferred it to the corresponding glottal filter.

The STRAIGHT features consisted of 21 aperiodicity energy bands and a spectral envelope represented as 40-dimensional MGC coefficients [50], extracted from 2048-point FFTs, a frequency warping factor of 0.42, and a power parameter of generalized cepstrum of 0. Apart from F0, the energy and spectral tilt were modified by mapping the first three Mel cepstrum coefficients (c_0, c_1 and c_2) of the MGC feature, and keeping the other coefficients unchanged (similar to [63]). In PML, the binary noise mask was 25-dimensional. The spectral envelope was extracted using STRAIGHT analysis and represented with exactly the same 40-dimensional MGC feature as in STRAIGHT. Again the same features as in STRAIGHT were mapped for PML.

2) DURATION MODIFICATION

The scaling ratios for the duration conversion of the voiced and unvoiced segments of speech were calculated as the mean ratio of the corresponding durations in the aligned segments from the two speaking-styles, measured across all (un)voiced segments in the data. These were found to be 1.08 and 0.88, i.e. the voiced and unvoiced regions were stretched and compressed, respectively (in line with the study of the duration of phonemic classes in Lombard speech in [29]). The durations were modified by applying cubic spline interpolation to the resulting feature time-series.

3) PRE-PROCESSING

In our earlier work on SSC [64], we modeled each feature separately for the sake of simplicity. However, since vocoder features are correlated, the MLM could potentially make use of these inter-relationships. Hence, feature concatenation was explored and found to improve results. A certain amount of contextual information in the feature domain could also be potentially useful to the mapping model. We explored three options for including contextual information: 1) deltas and delta-deltas of the concatenated features, 2) directly concatenating ± 1 adjacent frames (i.e. on either side of the current frame, with a total window size of 3 adjacent frames), and 3) concatenating ± 3 adjacent frames. By including contextual information, no noticeable differences were found when comparing the delta and delta-delta and ± 1 adjacent frames. Including ± 3 adjacent frames resulted in smoother feature contours, slightly better perceived quality (fewer distortions),

but a slightly lower Lombardness effect. Finally, we chose the ± 1 adjacent frames for the current study. Hence, the dimensionality of the mapped features was 36 for GlottDNN and 12 for PML and STRAIGHT.

Speaker-specific mean and variance normalization was applied to the features. This leads to a more balanced representation of the data in the feature space, such that the speaker-specific traits are averaged out and style-specific traits are retained to be modeled. Finally, in post mapping, the features were calculated using overlap-add over the adjacent frames with a flat window. The estimated features were then smoothed with median filtering using a window length of 30 for F0 and a window length of 3 for the MGC from STRAIGHT and PML and Gain from GlottDNN. In order to counter the potential differences in the total gains of the different vocoders in the listening tests, the utterance-level RMS values were normalized to that of the reference Lombard speech utterance.

4) MAPPING MODELS

The SGMMs were trained with K chosen from a range of 10 to 50 (in steps of 10) Gaussians, using 5-fold cross-validation. The EM algorithm was initialized randomly. For the BGMMs, K was set to 100, since BGMMs do not suffer from over-fitting with even a large number of Gaussians. (A separate 10-fold cross-validation test showed that the errors did not reduce significantly for larger values of K.) Furthermore, the BGMM component means and precisions were modeled with prior distribution $\mathcal{NW}(\boldsymbol{\mu}_0, \beta_0, \mathbf{W}_0, \nu_0)$, whose parameters were set similar to those recommended in [65]: $\boldsymbol{\mu}_0$ and $\mathbf{W}_0$ were set to the dataset mean and diagonal precision, $\beta_0 = 1$, and $\nu_0 = D + 2$. The concentration parameter $\boldsymbol{\alpha}_0$ was set to the all ones vector.

The feed-forward DNN [58] had 4 hidden layers with 250 hidden rectified linear unit (ReLu) neurons each. During optimization, the validation set was chosen as a random 20% sampling of the training set. The DNNs were optimized with RMSprop [58], [66] based on mean squared error and a mini-batch size of 100. Dropout regularization (15% of all the hidden layer units) and early stopping (maximum number of epochs and patience of 250 and 10, respectively, for the initial training, and 30 and 2 during model adaptation) were used to limit overfitting to the training data. DNNs were implemented using Keras (https://keras.io/).

VII. SYSTEM EVALUATION

The performance of the speaking style conversion system was evaluated using subjective listening tests comparing all 9 combinations of the 3 VOCs and 3 MLMs chosen in this study. Two separate tests were conducted to evaluate *degree of Lombardness* and *quality* of the converted speech samples, detailed below in Sections VII-A and VII-B respectively. Altogether 21 (11 male and 10 female) native talkers of Finnish took part in the listening tests, participating in both tests (with a short break in between the tests). Eleven of the listeners took the Lombardness test first and the rest started with the quality test. The listening tests were conducted in single-walled listening booths with a background noise level of less than 10 dB in the frequency range of the test samples using circumaural Sennheiser HD650 headphones. The tests were implemented using MATLAB's GUI (the Lombardness test system was adapted from [67]).

As one of the potential use cases, Lombard speech is desirable for the purpose of speech intelligibility enhancement research [30], [68], and therefore we also evaluated the intelligibility of the mapped speech. However, running subjective intelligibility tests for all the 9 different system configurations was not practically feasible. Instead intelligibility of our system was studied using a recently developed instrumental intelligibility metric called speech intelligibility in bits (SIIB, [69]). This measure is based on the mutual information between a clean reference and a noisy signal, and it performed well in a recent survey that compared several instrumental methods for measuring speech intelligibility [70]. In the current study, $\text{SIIB}^{\text{Gauss}}$ [70], a variation of SIIB that uses the information capacity of a Gaussian channel for mutual information calculation, was used. Babble noise was used to degrade the signals.

Finally, the proposed normal-to-Lombard conversion system was analyzed by i) visualizing the long term average spectra of the mapped utterances and by ii) objectively analyzing the distributions of the mapped key features F0, energy, and spectral tilt by comparing them to the same features from the natural normal and Lombard speech.

A. LOMBARDNESS TEST

This test was set up as a MUSHRA-like (MUltiple Stimuli with Hidden Reference and Anchor, [71]) test. Each trial aimed to evaluate the Lombardness of the mapped utterances from different VOC and MLM combinations of a single sentence (same speaker and linguistic content). In a single trial, the listeners were given reference samples consisting of the original utterance spoken in both normal and Lombard styles and a set of unlabeled samples with the same speaker and lexical content to be rated on a Lombardness scale from 0 to 100. The utterances to be rated included a set of mapped utterances and two hidden references of the original natural utterances in normal and Lombard styles (which were instructed to be rated as 0 and 100, respectively) to test the attentiveness of the listeners.

In standard MUSHRA, the listeners would have been asked to rate all nine possible VOC$\times$MLM combinations in each trial together with the two hidden references, which is too much to compare at once. Instead, the listeners rated a subset of four unique combinations in each trial, which were created by always sampling two out of three VOC and MLM options for the trial and presenting all their combinations. To include all possible pair-wise comparisons between different VOC and MLM variants for each test utterance, three different trials with the same utterance were required. As a result, the listeners were subjected to a total of 18 trials consisting of 6 different utterances (3 lexically unique sentences each

spoken by one male and female talker), with each trial having 2 reference utterances and 6 utterances to be rated (4 mapped and 2 hidden reference). The listeners were allowed to listen to the utterances as many times as they wished. The utterances and the two speakers were chosen randomly for the listening test after discarding the longest sentences (to minimize listening test duration) and excluding speakers that had produced very low perceived Lombardness as judged by the authors.

Before taking the test, the listeners were given a brief written description of Lombard speech, translated as: "*In Lombard speech, the speech becomes more pressed and energetic to remain intelligible in a noisy environment*". The listeners were also asked to focus on the style and try to ignore the speech quality. Each listener then had a training session in order to familiarize himself/herself with Lombard speech. In this training session, the listeners were able to listen to utterances in both styles. The utterances used in the training session were randomly chosen from speakers and sentences not used in either of the subjective tests. Furthermore, the listeners were asked to adjust the sound volume to a loud yet comfortable level during the training session, after which the volume was kept fixed for the duration of the actual test.

Since a standard paired t-test was not directly applicable to our data due to the deviation from standard MUSHRA, a repeated-measures ANOVA (RMANOVA) was carried out to test the main effects and interactions for VOCs and MLs while controlling for the effects of distributing the comparisons across three different trials, and also to see whether the utterance contents or speaker ID had any impact on the Lombardness ratings. The RMANOVA model therefore consisted of five factors: VOC$\times$MLM$\times$SEN$\times$SP$\times$REP, where SEN and SP represent the lexically unique sentence and speaker ID, while REP represents the number of times the SEN$\times$SP combination had been presented to the listener at the time of rating, adding up to a total of $3\times3\times3\times2\times3 = 162$ samples for each listener. The five missing VOC$\times$MLM combinations from each of the 18 trials were treated as missing data, resulting in 72 samples with actual ratings per subject. Within-subject correlations were modeled using a first-order autoregressive model and the reported degrees of freedom in results correspond to the corrected *df*s.

All main effects of the five factors and the interactions including MLM, VOC, or both, were tested, followed by pair-wise comparisons between the alternative MLM or VOC methods using Bonferroni-correction for unnormalized significance of $p < 0.05$ and using the RMANOVA model -based estimates of means for the comparisons. The mixed-model routine ("MIXED") of SPSS (v 24.0) was used to carry out the analyses due to its capability for properly handling missing data. Data from 2 male and 1 female listener were rejected from the analyses, as they had marked the hidden references incorrectly on more than one of the trials (indicating below 90% accuracy).

B. QUALITY TEST

The quality test was performed using the comparison category rating (CCR) test [72]. In this test, for a particular trial the listeners were presented with pairs of speech utterances. They were asked to rate the perceived quality of the second utterance in comparison to the first one using a continuous rating scale that translates to English as: -3, much worse; -2, worse; -1, slightly worse; 0, almost similar; 1, slightly better, 2, better; 3, much better. The listeners could listen to the utterance pairs as many times as they wished. In a single trial, each utterance pair consisted of a mapped utterance and its corresponding natural Lombard utterance. This was presented in both orders and also including null pairs where both utterances in the pair were from the natural Lombard reference (expected to be rated as 0; in order to test the attentiveness of the listeners). For each utterance, there were 19 trials (including both CCR utterance pair orders for the 9 VOC$\times$MLM combinations and 1 null pair). Ideally, it would have been preferable to include all 36 possible pairs from the 9 VOC$\times$MLM combinations to better compare them directly against each other, but this would have made the test unfeasible due to the large number of trials. Similarly to the Lombardness test, we used 6 utterances, consisting of 3 lexically unique sentences spoken by one talker of both genders (randomly chosen and different from the Lombardness test and the training sessions). The listeners were exposed to a total of 114 trials, and were allowed to take a short break after rating the first 57 CCR utterance pairs.

Prior to the actual test, the listeners had a training session where they had to rate 4 CCR utterance pairs (including a null pair). These training utterances were randomly chosen from the same set that was used for the training session in the Lombardness test. As in the Lombardness test, the listeners were asked to adjust the volume during the training session, after which it was unchangeable.

The average of the scores for each unique utterance pair from the CCR test, i.e. from the pairs of the utterances presented in both orders, known as the comparison mean opinion score (CMOS) [72], was first calculated. The corresponding marginal means for each VOC$\times$MLM combination was then derived from these. Hence, the CMOS score in our current study corresponds to how much better the reference natural Lombard utterance is in comparison to the mapped Lombard. Therefore, a lower CMOS score indicates higher quality. A male listener was rejected from the analyses since he had marked all scores as either $+3$ or -3 which did not carry any information with respect to the relative quality of different methods. The rest of the listeners marked the null pairs with a score of close to 0 and were hence considered for analysis. As noted in [73], there is a significant listener specific bias in the CMOS scores. To counter this, we mean-normalized the CMOS scores across each listener (to zero mean across all comparisons) before aggregating the results across the different system variants. Since variance of the responses was not normalized, this approach allows us to study the magnitude of

differences in quality ratings for different system variants. All VOCs, MLMs and their combinations were then compared using the two-sided Mann-Whitney U-test (as in [73]).

VIII. RESULTS

A. LOMBARDNESS TEST

The main results from the Lombardness listening test are shown in Figure 3, where the horizontal lines and asterisks indicate significant differences as obtained from the mixed effects model for the VOC and ML comparison and from the two sided pairwise t-test for the VOC×ML comparison. The statistical analyses indicate that there were significant main effects of VOC ($F(1064.53) = 73.86$, $p < 0.001$) and ML ($F(500.16) = 4.73$, $p = 0.009$), and significant interactions of VOC×MLM ($F(1095.21) = 4.34$, $p = 0.002$) and VOC*REP ($F(906.76) = 3.97$, $p = 0.003$). The main effects of utterance, repetition, or speaker ID, or any other interactions between the factors were not significant. This means that the ratings were affected by the choice of a vocoder and a mapping method, but not by the linguistic contents or the speaker. The number of times the same utterance had been heard earlier had a slight impact on the relative ratings of different vocoders (see below).

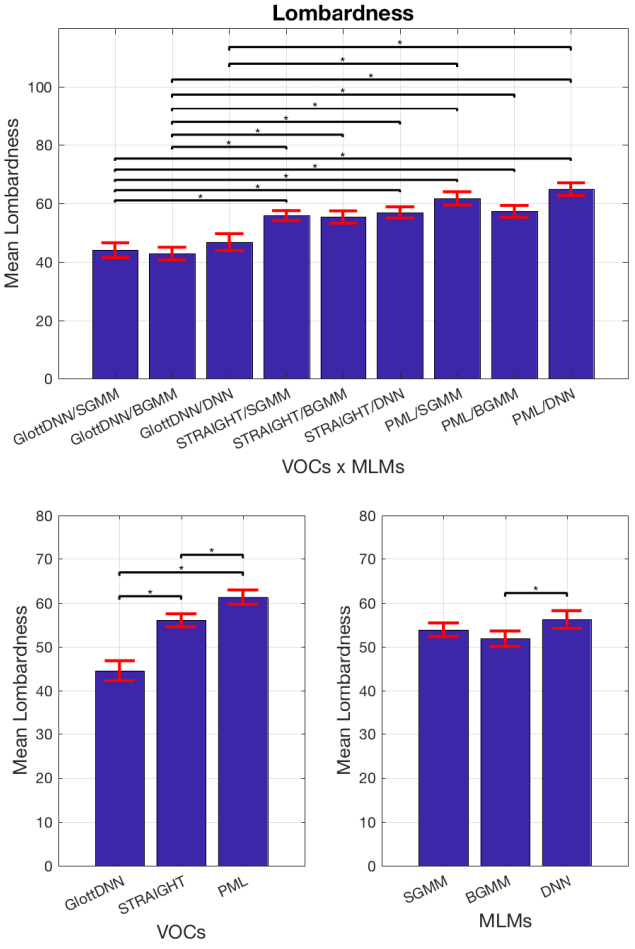

FIGURE 3. The mean Lombardness scores over all the 9 VOC×ML combinations (top) and separately for each VOC and MLM (bottom). Standard errors across all subjects are shown in red. The horizontal lines and asterisks indicate the significant differences ($p < 0.05$ after Bonferroni correction for multiple comparisons). A score of 0 and 100 correspond to the mapped utterances having as much Lombardness as the reference normal and Lombard utterances respectively.

Comparing the RMANOVA model-based means of Lombardness scores of the VOCs, all differences between the vocoders are significant with PML having higher average Lombardness rating ($M = 61.22$, standard error $SE = 2.92$) than GlottDNN ($M = 43.48$, $SE = 2.93$; $p < 0.001$) and STRAIGHT ($M = 57.14$, $SE = 2.91$; $p = 0.008$), and STRAIGHT is also rated higher than GlottDNN ($p < 0.001$). As for the MLMs, DNN has a significantly higher average Lombardness ($M = 59.21$, $SE = 3.28$) than BGMM ($M = 49.39$, $SE = 3.42$; $p = 0.008$), whereas there is no significant difference between SGMM ($M = 53.24$, $SE = 3.33$) and DNN or between SGMM and BGMM.

As for the interaction between vocoders and repetitions, pairwise comparisons revealed that the Lombardness difference between STRAIGHT and PML increased gradually across the three repetitions of the same utterance, reaching significance only at the third repetition (as measured with a highly conservative Bonferroni correction). Closer analysis revealed that the Lombardness-ratings of STRAIGHT decreased systematically from the first repetition (mean 62.73) to the third repetition of the same utterance (mean 52.02) while GlottDNN and PML were more consistent across all repetitions (GlottDNN having 45.15 on the first repetition and 45.60 on the last repetition while the corresponding numbers for PML are 63.50 and 60.50, respectively) even though the presentation and comparison orders were fully randomized across all participants. The reason for this finding is currently unclear, although the result suggests that the listeners somehow started to consider STRAIGHT as less Lombard-like the more they heard the vocoder with the same utterance. On the other hand, the first impressions of the Lombardness in STRAIGHT-vocoded samples were comparable to those of PML on every new utterance.

Finally, we carried out exhaustive pair-wise comparisons between all the 9 combinations of MLMs and VOCs using two-tailed t-tests for the subject mean ratings with Bonferroni corrected significance levels ($p < 0.05$). The significant comparisons are: STRAIGHT/SGMM ($M = -55.82$, $SE = -1.69$) is better than GlottDNN/SGMM ($M = -44.03$, $SE = -2.52$; $p = -0.016$ and $df = 18$ for all t-tests) and better than GlottDNN/BGMM ($M = -42.83$, $SE = -2.17$; $p = -0.001$); STRAIGHT/BGMM ($M = -55.33$, $SE = -2.09$) is better than GlottDNN/BGMM ($p = -0.008$); STRAIGHT/DNN ($M = -56.87$, $SE = -1.95$) is better than GlottDNN/SGMM ($p = -0.011$) and better than GlottDNN/BGMM ($p = -0.001$); PML/SGMM ($M = -61.68$, $SE = -2.27$) is better than GlottDNN/SGMM ($p < 0.001$), better than GlottDNN/BGMM ($p < 0.001$), and better than GlottDNN/DNN ($M = -46.77$, $SE = -2.87$; $p = -0.009$); PML/BGMM ($M = -57.29$, $SE = -1.99$) is better than GlottDNN/SGMM ($p = -0.008$), and better than GlottDNN/BGMM ($p = -0.001$);

and finally PML/DNN (M = −64.87, SE = −2.13) is better than GlottDNN/SGMM (p < 0.001), better than GlottDNN/BGMM (p < 0.001), and better than GlottDNN/DNN (p = −0.001). The top panel in Figure 3 shows the results for all 9 VOC×ML combinations.

B. QUALITY TEST

Figure 4 shows the results of the average and standard errors of mean-normalized CMOS scores from the VOC, MLM, and VOC×MLM categories. The lines and asterisks indicate the significant differences (Bonferroni corrected with $p < 0.05$ and obtained from the two-sided Mann-Whitney U-test). The overall mean CMOS score over all the methods was 1.9.

Comparing the aggregated mean-normalized CMOS scores of the VOCs, the following significant results were observed (the values are calculated as the difference in the average mean-normalized CMOS scores): GlottDNN (M = −0.064, SE = 0.02) is better than PML (M = 0.06, SE = 0.01; p < 0.001) and STRAIGHT (M = 0.00, SE = 0.02) is better than PML (p = 0.043). There is no significant difference between GlottDNN and STRAIGHT.

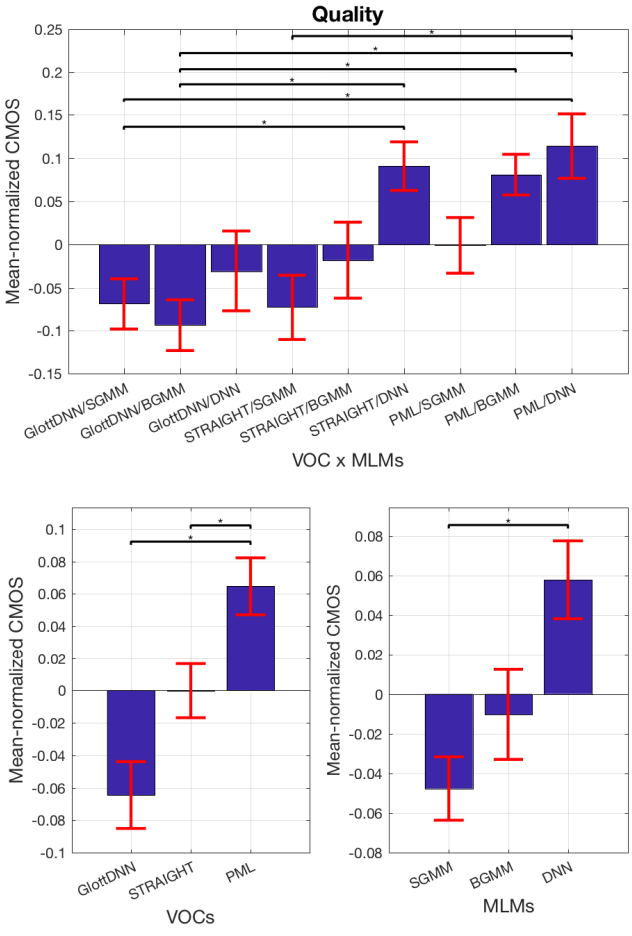

FIGURE 4. Bar plots indicated the mean of the listener specific mean-normalized CMOS scores over all the 9 VOC×ML combinations, 3 VOCs, and 3 MLMs (lower is better). Standard errors are shown in red. The horizontal lines and asterisks indicate the significant differences (Bonferroni corrected for unnormalized $p < 0.05$).

Similarly comparing the results for the MLMs, only one significant result was found: SGMM (M = −0.05, SE = 0.01) is better than DNN (M = 0.06, SE = 0.02; p = 0.001). BGMM (M = −0.01, SE = 0.02) does not have any significant differences from the other MLMs.

Finally, comparing the results for all 9 VOC×MLM, the following significant results were found: GlottDNN/SGMM (M = −0.07, SE = 0.03) is better than STRAIGHT/DNN (M = 0.09, SE = 0.03; p = 0.049) and better than PML/DNN (M = 0.11, SE = 0.04; p = 0.040); GlottDNN/BGMM (M = −0.09, SE = 0.03) is better than STRAIGHT/DNN (p = 0.009), better than PML/BGMM (M = 0.08, SE = 0.02; p = 0.004), and better than PML/DNN (p = 0.006); and finally STRAIGHT/SGMM (M = −0.07, SE = 0.04) is better than PML/DNN (p = 0.044). GlottDNN/DNN (M = −0.03, SE = 0.05), STRAIGHT/BGMM (M = −0.02, SE = 0.04) and PML/SGMM (M = 0.00, SE = 0.03) were not significantly different from any other methods.

C. INSTRUMENTAL INTELLIGIBILITY TEST

We first verified that the SIIB produces meaningful results for natural and copy synthesized normal and Lombard speech with different vocoders (i.e., vocoding and synthesis without mapping) at different SNRs. As can be observed from Fig. 5, copy synthesis with STRAIGHT maintained intelligibility comparable to natural utterances. However, intelligibility scores of PML and even more so of GlottDNN do not reach the scores of the natural utterances, and this difference also holds for very high SNRs (not shown separately). Although the reason for this discrepancy is unclear, one potential reason could be that $\text{SIIB}^{\text{Gauss}}$ metric favours the prominent harmonic structure in STRAIGHT-vocoded speech caused by the vocoder's impulse-type excitation waveform. Compared to STRAIGHT, PML and GlottDNN use more complicated excitation waveforms which results in less prominent harmonics in synthesized speech. Even though this feature has been found to reduce "buzziness" in speech synthesis [38], the use of more complicated excitation waveforms in PML and GlottDNN get penalized by $\text{SIIB}^{\text{Gauss}}$ for some reason (see also Section VIII-D).

Since $\text{SIIB}^{\text{Gauss}}$ does not treat different vocoders in the same manner in the copy synthesis conditions, it is not possible to directly compare different VOCs against each other in the mapping process. It is possible, however, to analyze different MLMs by using a selected vocoder and by comparing the SIIB scores of converted speech to those computed from copy synthesized utterances generated with the same VOC from natural normal and Lombard speech. These results are shown in Figure 6 in terms of the intelligibility gains (bits/s) in noise for different system configurations when compared to copy-synthesized normal speech. As can be observed, feature mapping improves the $\text{SIIB}^{\text{Gauss}}$ score with reference to the copy synthesis normal style utterance using that same VOC (except for GlottDNN with SGMM and VBGMM at 5 dB SNR). Increase in estimated intelligibility is higher for STRAIGHT

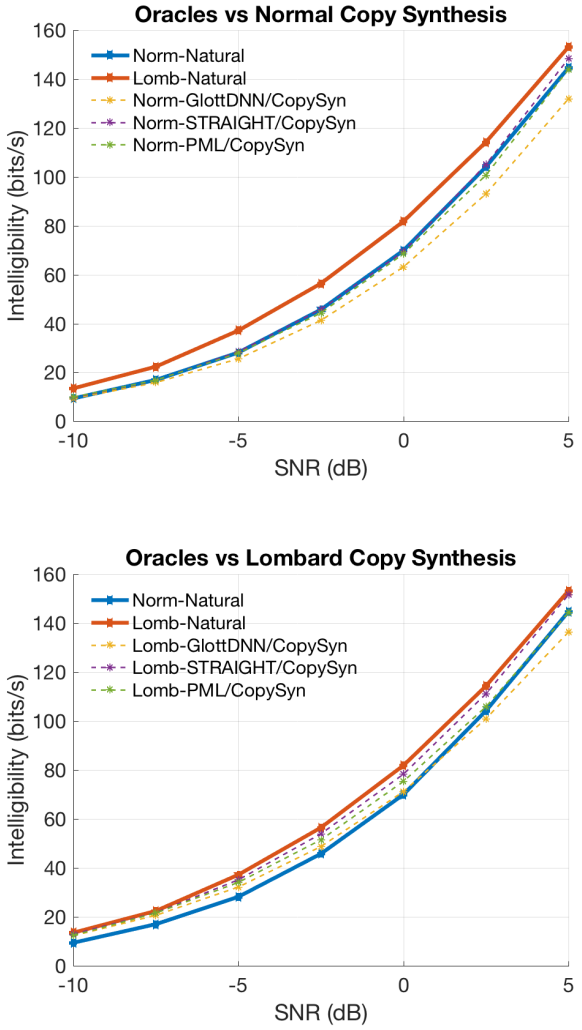

FIGURE 5. SIIB$^{\text{Gauss}}$ intelligibility scores at different SNRs, comparing the natural utterances and the copy synthesized utterances generated with GlottDNN, STRAIGHT and PML. Top panel: copy synthesis of normal utterances. Bottom panel: copy synthesis of Lombard utterances.

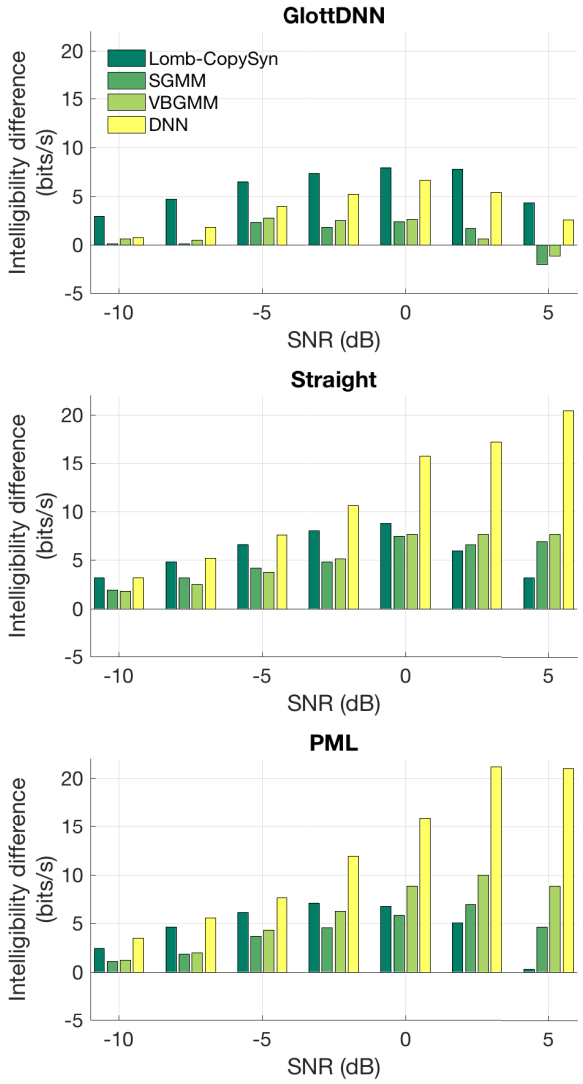

FIGURE 6. The improvement in SIIB$^{\text{Gauss}}$ intelligibility scores at different SNRs compared to the copy synthesized normal style utterance in the same noise conditions. For each VOC, intelligibility after mapping with each of the different MLMs is shown. Intelligibility gains for copy synthesized Lombard utterances are shown as a reference.

and PML compared to GlottDNN. Comparing the MLMs, the DNN leads to the highest intelligibility gain among the MLMs. The SGMM and VBGMM have similar intelligibility gains. These observations are in line with the results of the subjective Lombardness test in Section VIII-A, thereby also confirming that the subjective ratings of Lombardness by our subjects are closely related to instrumental measure of speech intelligibility.

D. FEATURE ANALYSIS

Figure 7 shows the long term average spectra (LTAS) of the natural utterances in both styles along with the copy synthesis utterances for each VOC. Utterances mapped with the DNN—the method that was found to produce highest Lombardness in both subjective listening and instrumental intelligibility tests—are also plotted for each VOC. It can be seen that the LTAS of the copy synthesis utterances in both styles follow those of the natural utterances. This suggests that the differences in SIIB$^{\text{Gauss}}$ among the VOCs do not arise from inherent differences in the spectral shape of the VOC outputs. The third plot in Figure 7, corresponding to LTAS after the normal-to-Lombard mapping, shows that PML is closest to the target natural Lombard spectrum, followed by STRAIGHT. In contrast, GlottDNN deviates from the other two with a much smaller effect of the mapping, especially at <2 kHz where the spectrum remains largely unmodified from normal speaking style. In total, the average spectra are in line with the results observed in the subjective Lombardness (Section VIII-A) and objective intelligibillity (Section VIII-C) tests where the DNN had the highest Lombardness and SIIB$^{\text{Gauss}}$ scores respectively. They also reveal that the mapping of the GlottDNN parameters has not been as successful as for the other two vocoders, either due to the higher dimensionality of the source features in GlottDNN,

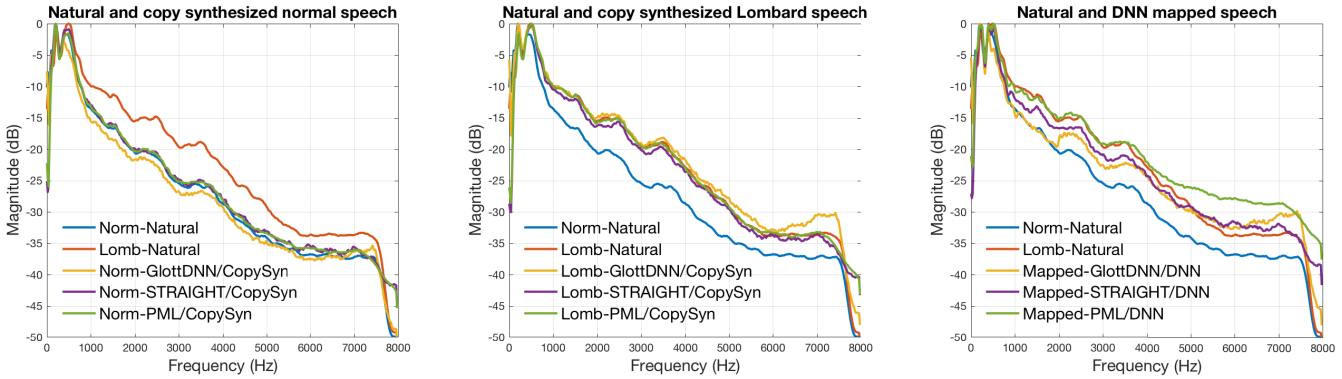

FIGURE 7. Long term average spectra (LTAS) in normal and Lombard styles. LTAS of natural normal and natural Lombard speech are shown in all panels. Panels from left to right show (a) LTAS of copy synthesis normal speech generated with three VOCs, (b) LTAS of copy synthesis Lombard speech generated with three VOCs, and (c) LTAS of DNN-mapped speech.

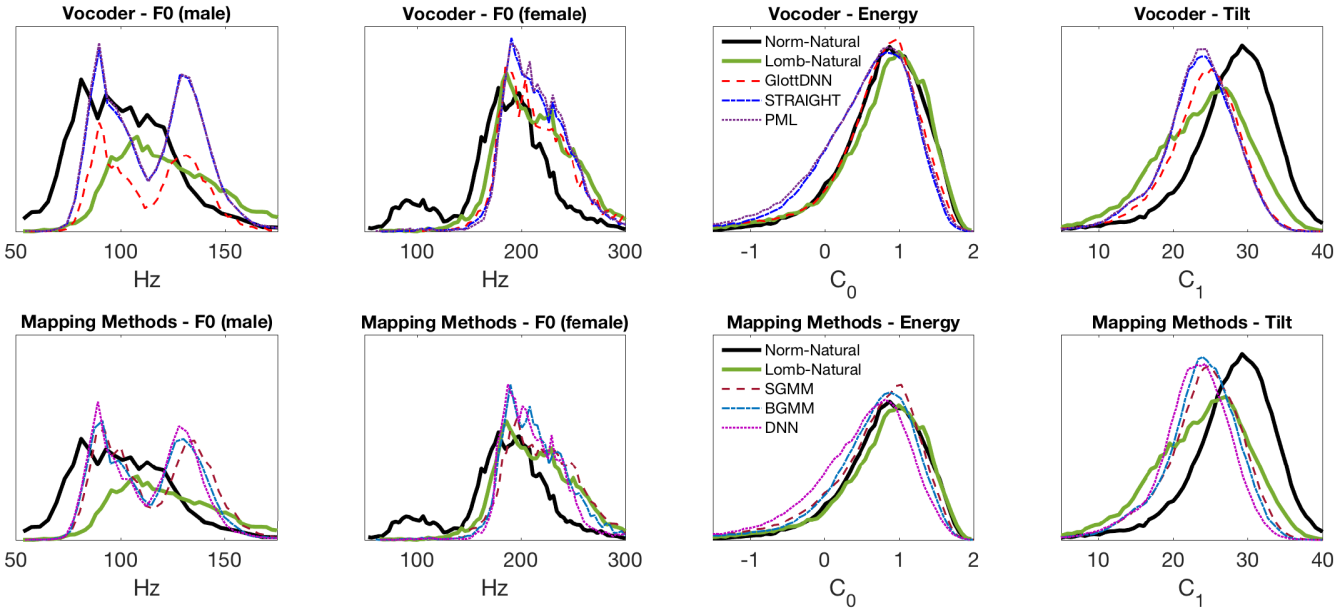

FIGURE 8. Histogram plots showing the distribution of the key features mapped in this study: F0, frame-wise energy (c_0), and spectral tilt (c_1) (calculated using the SPTK toolkit [61]). The black and green solid lines show the feature distributions of the reference normal and reference Lombard speech data, respectively. The rest of the colored and dashed lines show the distributions of these features after the mapping with different methods. The distributions for different VOCs are obtained by averaging across all MLMs (top row), and distributions for MLMs are obtained by averaging across all VOCs (bottom row).

or due the distributional properties of the source features (LSFs).

Finally, the distributions of the mapped features (i.e. F0, frame-wise energy, and spectral tilt from the voiced frames) were also analyzed using the overall histogram distributions of the features corresponding to the natural normal, natural Lombard and mapped Lombard utterances. The F0 was extracted similarly to all vocoders using the SPTK toolkit [61]. The frame-wise energy and spectral tilt features were calculated as the zeroth (c_0) and first (c_1) standard MFCC coefficient, respectively, and independently of any vocoder-specific feature extractors. The histograms were calculated using the same 100 bins for all VOC and MLM variants and were normalized to sum up to 1. The resulting histogram distributions are shown in Figure 8, where F0 is shown separately for male and female talkers. Ideally, the distributions of the mapped features should be far away from the distribution of the natural normal and close to those of the natural Lombard. This trend is clearly visible in the mapped features.

For a quantitative analysis of the feature distributions, the net divergence, d_{net}, measures in Table 1 show the relative

TABLE 1. Net Divergence, d_{net}, of the feature distributions calculated as shown in 9. The features shown are F0, energy (C_0), and spectral tilt (C_1).

Feats	VOCs			MLMs		
	Glott-DNN	STRA-IGHT	PML	SGMM	BGMM	DNN
F0-M	**0.033**	0.032	0.030	**0.054**	0.025	0.014
F0-F	0.084	0.084	**0.087**	**0.099**	0.081	0.071
C_0	**-0.0005**	-0.002	-0.0006	**0.0007**	-0.002	-0.002
C_1	0.075	0.092	**0.093**	0.070	0.083	**0.107**

distance of the feature distributions to the natural normal and natural Lombard speech. d_{net} is calculated with the Jenson-Shannon divergence (JSD), the symmetric version of the Kullback-Leibler divergence (KLD) as

$$\begin{aligned} d_{net} &= JSD(\text{mapped}||\text{natural normal}) \\ &\quad - JSD(\text{mapped}||\text{natural Lombard}) \\ \text{where,}\quad JSD(p||q) &= \frac{1}{2}KLD(p||\frac{p+q}{2}) + \frac{1}{2}KLD(q||\frac{p+q}{2}) \end{aligned} \tag{9}$$

where a higher d_{net} means that the features of the mapped utterances are, on average, further away from those of the natural normal and closer to the Lombard speech. d_{net} values above 0 indicate that the mapped feature distribution is closer to the natural Lombard speech than normal speech, which is seen for all the features except frame-wise energy. As a reference for the scale of the d_{net} values, the values of JSD(natural normal||natural Lombard), JSD(mapped||natural normal) and JSD(mapped||natural Lombard) averaged over all the features are 0.039, 0.054 and 0.018 respectively. An interesting finding here is that the divergence between natural normal and natural Lombard speech is smaller than between the natural normal and the mapped utterances, i.e. some of the features in the mapped utterances are further away from normal speech than those of natural Lombard. This is in fact desired, as several of the original speakers had rather mild Lombard effect and model adaptation to a subset of speakers was carried out to increase the Lombardness effect (see Section VI-B). However, it is difficult to accurately summarize the changes on a one-dimensional continuum due to the non-normality of the distributions.

Among the VOCs, GlottDNN has the highest net divergence in terms of male F0 and frame-wise energy, whereas PML has high separability for the female F0 and spectral tilt. Comparing the MLMs, DNN has the highest separation for spectral tilt, whereas the use of SGMM leads to the largest divergence in the rest of the features. These results are roughly in line with the results of the subjective Lombardness test from Section VIII-A, where DNN, SGMM, and PML had the highest levels of perceived Lombardness.

IX. DISCUSSION AND CONCLUSION

This work described a general parametric system that is capable of achieving style conversion of speech utterances. The specific case of a vocal effort based style conversion, normal-to-Lombard, was detailed for a scenario with limited parallel Finnish data. The system consists of vocoders for feature extraction and speech synthesis, and machine learning models with parallel learning for mapping the selected features. Three well-known vocoders and mapping methods were compared in the experiments. The system was tested with two subjective tests measuring Lombardness and quality and with an instrumental speech intelligibility metric. It was noted that all methods achieve a significant shift in the speaking style towards Lombard in the speech utterances with some degradation in the perceived quality.

The results from the listening tests and the instrumental intelligibility tests show that DNNs created the largest Lombardness effect, although at the cost of perceived signal quality. In addition, SGMMs and BGMMs performed very similarly. This is an indication that the BGMMs are capable of handling overfitting without the expensive hyper-parameter tuning that is required for SGMMs (e.g., the 5-fold cross validation used here; see Section VI-C4). Comparing the different vocoders, it was observed that GlottDNN is one of the best performing VOCs in terms of perceived quality but also leads to the smallest Lombardness score. This effect is also observable in the instrumental intelligibility tests where the GlottDNN has the smallest improvement in comparison to the natural normal reference. STRAIGHT has a good balance between these two measures, and PML is the best in terms of Lombardness and the worst in terms of perceived quality. The reduced Lombardness effect of the utterances mapped with GlottDNN could be because of feature dimensionality (36 mapped features for GlottDNN and 12 for the other two VOCs) combined with MLMs being designed to avoid overfitting, where estimation of feature covariances (explicitly in GMMs or implicitly in DNNs) is difficult due to limited training data. Another possibility is that the high-dimensional parametrization of the glottal source, when mapped together with F0 and energy, constrains the shift in the Lombardness, as not all aspects of the source undergo as substantial a change between styles as the mere tilt parameters (c_1 and c_2 MGC coefficients of spectral envelope) mapped in the other two VOCs. Overall, PML with SGMM mapping appears as one of the best compromises, with a high Lombardness score and a comparatively low quality degradation. The findings of the Lombardness listening tests were supported by the feature analysis, which shows that the features of the normal speech are indeed being mapped closer to the Lombard reference.

From the above results, it is clear that there is an inverse relationship between how well each of the methods performs in terms of Lombardness and perceived quality, suggesting that there is currently a trade-off between these two evaluating parameters. If the current parametric SSC system were to be used in a real-world application, this trade-off would limit the style conversion effect we are capable of achieving in the target style (currently Lombard) while still maintaining an acceptable level of quality. However, such a trade-off should not be an inherent property of the SSC problem, but is likely related to our limited training data and shortcomings in the current technical solutions and could be therefore addressed in the future work.

Considering this, one direction for future research could be working with non-parallel learning schemes, which will allow using considerably more training data from conversational recordings in natural settings that may have different speaking styles along the vocal effort continuum. This would allow either inclusion of more features to be mapped between the speaking styles (e.g., vocal tract parametrizations as well), more precise context- or content-dependent mapping of the current relatively small set of key features, or all of these at

the same time. Another approach to work around the data scarcity problem of the current SSC system would be the use of components pre-trained on large corpora, and then adapting these for different styles with the data available. For instance, a WaveNet with pre-trained weights could have been potentially used as a vocoder in the current system, but an open source implementation was not available at the time of experimentation and was hence not explored in the current tests. It would be interesting to see how the WaveNet compares with the standard parametric vocoders in this setting in future experiments. Once the basic recipe for high Lombardness and high-quality SSC has been established, a future study should investigate how the Lombardness scores from the subjective listening tests and the SIIB scores from the objective intelligibility evaluation actually translate in subjective speech intelligibility in different noise conditions. The implications of the current and future systems of SSC on different styles along the vocal effort continuum should be also explored. In this respect, the present study should be taken as the first steps towards this direction.

ACKNOWLEDGMENT

The authors would also like to thank all the participants of the listening tests.

REFERENCES

[1] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *J. Speech, Lang., Hearing Res.*, vol. 14, pp. 677–709, Dec. 1971.
[2] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech, Lang., Hearing Res.*, vol. 29, no. 4, pp. 434–446, Dec. 1986.
[3] R. M. Uchanski, S. S. Choi, L. D. Braida, C. M. Reed, and N. I. Durlach, "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate," *J. Speech, Lang., Hearing Res.*, vol. 39, no. 3, pp. 494–509, Jun. 1996.
[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.
[5] Y. Stylianou, "Voice transformation: A survey," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3585–3588.
[6] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.
[7] *One-Way Transmission Time*, document ITU-T Rec. G.114, International Telecommunication Union, Geneva, Switzerland, May 2003.
[8] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Commun.*, vol. 51, no. 3, pp. 268–283, Mar. 2009.
[9] D. Erro, E. Navas, I. Hernáez, and I. Saratxaga, "Emotion conversion based on prosodic unit selection," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 974–983, Jul. 2010.
[10] F. Tesser, E. Zovato, M. Nicolao, and P. Cosi, "Two vocoder techniques for neutral to emotional timbre conversion," in *Proc. 7th ISCA Workshop Speech Synth.*, Kyoto, Japan, Sep. 2010, pp. 130–135.
[11] M. Wang, M. Wen, K. Hirose, and N. Minematsu, "Emotional voice conversion for mandarin using tone nucleus model—Small corpus and high efficiency," in *Proc. Speech Prosody*, Kyoto, Japan, May 2012, pp. 163–166.
[12] J. Latorre, V. Wan, and K. Yanagisawa, "Voice expression conversion with factorised HMM-TTS models," in *Proc. Interspeech*, Singapore, Sep. 2014, pp. 1514–1518.
[13] A. K. Vuppala and S. R. Kadiri, "Neutral to anger speech conversion using non-uniform duration modification," in *Proc. ICIIS*, Gwalior, India, Dec. 2014, pp. 1–4.
[14] S. Vekkot and S. Tripathi, "Vocal emotion conversion using WSOLA and linear prediction," in *Proc. SPECOM*, Leipzig, Germany, 2017, pp. 777–787.
[15] Z. Inanoglu and S. Young, "A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 490–493.
[16] M. Koutsogiannaki, P. N. Petkov, and Y. Stylianou, "Intelligibility enhancement of casual speech for reverberant environments inspired by clear speech properties," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 65–69.
[17] M. D. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Commun.*, vol. 48, no. 5, pp. 549–558, May 2006.
[18] H. Konno, M. Kudo, H. Imai, and M. Sugimoto, "Whisper to normal speech conversion using pitch estimated from spectrum," *Speech Commun.*, vol. 83, pp. 10–20, Oct. 2016.
[19] Z. Tao, X.-D. Tan, T. Han, J.-H. Gu, Y.-S. Xu, and H.-M. Zhao, "Reconstruction of normal speech from whispered speech based on RBF neural network," in *Proc. IITSI*, Jinggangshan, China, Apr. 2010, pp. 374–377.
[20] G. N. Meenakshi and P. K. Ghosh, "Whispered speech to neutral speech conversion using bidirectional LSTMs," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 491–495.
[21] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad, "Fundamental frequency generation for whisper-to-audible speech conversion," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 2579–2583.
[22] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Med. Eng. Phys.*, vol. 24, nos. 7–8, pp. 515–520, Sep./Oct. 2002.
[23] K. Nathwani, G. Richard, B. David, P. Prablanc, and V. Roussarie, "Speech intelligibility improvement in car noise environment by voice transformation," *Speech Commun.*, vol. 91, pp. 17–27, Jul. 2017.
[24] A. Calzada and J. C. Socoró, "Vocal effort modification through harmonics plus noise model representation," in *Proc. NOLISP*, Las Palmas, Spain, Nov. 2011, pp. 96–103.
[25] D.-Y. Huang, S. Rahardja, and E. P. Ong, "Lombard effect mimicking," in *Proc. SSW*, Kyoto, Japan, Sep. 2010, pp. 258–263.
[26] K. I. Nordstrom, G. Tzanetakis, and P. F. Driessen, "Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 6, pp. 1087–1096, Aug. 2008.
[27] C. d'Alessandro and B. Doval, "Experiments in voice quality modification of natural speech signals: The spectral approach," in *Proc. ESCA/COCOSDA Workshop Speech Synth.*, Blue Mountains, NSW, Australia, Nov. 1998, pp. 277–282.
[28] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, 2010, pp. 1–6.
[29] 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. Audio, Speech, Language Process.*, vol. 17, no. 2, pp. 366–378, Feb. 2009.
[30] 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 Commun.*, vol. 55, no. 4, pp. 572–585, May 2013.
[31] Y. Wang *et al.* (2018). "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis." [Online]. Available: https://arxiv.org/abs/1803.09017
[32] J. Shen *et al.*, "Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.
[33] S. O. Arik *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proc. ICML*, 2017, pp. 1–17.
[34] J. Vít, Z. Hanzlíček, and J. Matoušek, "On the analysis of training data for WaveNet-based speech synthesis," in *Proc. ICASSP*, 2018, pp. 5684–5688.
[35] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, Dec. 2017, pp. 712–718.
[36] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN-A full-band glottal vocoder for statistical parametric speech synthesis," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 2473–2477.
[37] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, nos. 3–4, pp. 187–207, Apr. 1999.

[38] G. Degottex, P. Lanchantin, and M. Gales, "A log domain pulse model for parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 57–70, Jan. 2018.
[39] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B, Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.
[40] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.
[41] 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. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 2771–2775.
[42] T. Raitio *et al.*, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.
[43] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.
[44] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 184–194, Apr. 2014.
[45] Q. Hu, Y. Stylianou, R. Maia, K. Richmond, J. Yamagishi, and J. Latorre, "An investigation of the application of dynamic sinusoidal models to statistical parametric speech synthesis," in *Proc. Interspeech*, Singapore, Sep. 2014, pp. 780–784.
[46] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Upper Saddle River, NJ, USA: Prentice-Hall, 1978.
[47] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.
[48] 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. ICASSP*, Shanghai, China, Mar. 2016, pp. 5120–5124.
[49] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. MAVEBA*, 2001, pp. 59–64.
[50] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis—A unified approach to speech spectral estimation," in *Proc. ICSLP*, Yokohama, Japan, Sep. 1994, pp. 1043–1046.
[51] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.
[52] E. Jokinen, U. Remes, M. Takanen, K. Palomäki, M. Kurimo, and P. Alku, "Spectral tilt modelling with GMMs for intelligibility enhancement of narrowband telephone speech," in *Proc. Interspeech*, Singapore, Sep. 2014, pp. 2036–2040.
[53] E. Jokinen, U. Remes, and P. Alku, "Comparison of Gaussian process regression and Gaussian mixture models in spectral tilt modelling for intelligibility enhancement of telephone speech," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 85–89.
[54] L. Li, Y. Nankaku, and K. Tokuda, "A Bayesian approach to voice conversion based on GMMs using multiple model structures," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 661–664.
[55] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, Seattle, WA, USA, 1998, pp. 285–288.
[56] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, 2000.
[57] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep., Oct. 2007.
[58] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[59] D. Ellis. (2003). *Dynamic Time Warp (DTW) in MATLAB*. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/
[60] *Speech and Multimedia Transmission Quality (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database, Version 1.2.4*, ETSI, Sophia Antipolis, France, 2011.
[61] SPTK Working Group. (2014). *Speech Signal Processing Toolkit (SPTK) Version 3.8*. [Online]. Available: http://sp-tk.sourceforge.net/
[62] D. Talkin. (2015). *REAPER: Robust Epoch and Pitch EstimatoR*. [Online]. Available: https://github.com/google/REAPER
[63] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise," in *Proc. Interspeech*, Portland, OR, USA, Sep. 2012, pp. 631–634.
[64] A. R. López, S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Speaking style conversion from normal to Lombard speech using a glottal vocoder and Bayesian GMMs," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1363–1367.
[65] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
[66] G. Hinton, N. Srivastava, and K. Swersky. (2012). *Neural Networks for Machine Learning—Lecture 6a Overview of Mini-Batch Gradient Descent*. Accessed: Mar. 20, 2018. [Online]. Available: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf
[67] E. Vincent. (2005). *Mushram 1.0*. Accessed: Mar. 20, 2018. [Online]. Available: http://c4dm.eecs.qmul.ac.uk/downloads/
[68] E. Jokinen, U. Remes, and P. Alku, "Intelligibility enhancement of telephone speech using Gaussian process regression for normal-to-Lombard spectral tilt conversion," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1985–1996, Oct. 2017.
[69] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.
[70] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2153–2166, Nov. 2018.
[71] *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*, document Rec. ITU-R BS.1534-3, International Telecommunication Union, Geneva, Switzerland, Nov. 2015.
[72] *Methods for Objective and Subjective Assessment of Quality*, document Rec. ITU-R P.800, International Telecommunication Union, Aug. 1996.
[73] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 3976–3980.

SHREYAS SESHADRI was born in 1991. He received a dual M.Sc. degree in research on information and communication technologies from the Université Catholique de Louvain, Belgium, and the Universitat Politècnica de Catalunya, Spain, in 2014. From 2011 to 2012, he was a Research Intern with the Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany. He is currently pursuing the D.Sc. (Tech.) degree with the Department of Signal Processing and Acoustics, Aalto University, Finland. His research interests include machine learning and speech processing. He was a recipient of the Category-A Erasmus Mundus Scholarship for the same.

LAURI JUVELA was born in 1989. He received the M.Sc. (Tech.) from Aalto University, Finland, in 2015, where he is currently pursuing the D.Sc. degree in speech and language technology. From 2015 to 2016, he was a Visiting Researcher with the National Institute of Informatics, Tokyo, Japan. His research interests include speech signal processing, text-to-speech, and deep learning, with a focus on generative methods.

OKKO RÄSÄNEN was born in Finland, in 1984. He received the M.Sc. (Tech.) degree in language technology from the Helsinki University of Technology, Finland, in 2007, and the D.Sc. (Tech.) degree in language technology from Aalto University, Finland, in 2013. He also holds the Title of Docent (Adjunct Professor) from Aalto University in the area of spoken language processing.

He is currently an Assistant Professor with the Laboratory of Signal Processing, Tampere University of Technology, Finland, and a part-time Visiting Researcher with Aalto University. In 2015, he was a Visiting Researcher with the Language and Cognition Lab, Stanford University, Stanford, CA, USA. His research interests include computational modeling of language acquisition, cognitive aspects of language processing, context-aware computing, multimodal data analysis, and speech processing in general. He is a member of the International Speech Communication Association and the Cognitive Science Society.

PAAVO ALKU received the M.Sc., Lic.Tech., and Dr.Sc.(Tech) degrees from the Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an Assistant Professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993, and also with the University of Turku, Finland, from 1994 to 1999. He is currently a Professor of speech communication technology with Aalto University, Espoo. His research interests include analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modeling of speech, speech enhancement, and cerebral processing of speech.