Ana Ramírez López

# Methods for speaking style conversion from normal speech to high vocal effort speech

**School of Electrical Engineering**

Thesis submitted for examination for the degree of Licentiate of Science in Technology.

Espoo 18.08.2020

**Thesis supervisor and advisor:**

Prof. Paavo Alku

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
LICENTIATE'S THESIS

Author: Ana Ramírez López

Title: Methods for speaking style conversion from normal speech to high vocal effort speech

Date: 18.08.2020 Language: English Number of pages: 11+71

Department of Signal Processing and Acoustics

Professorship: Speech Communication Technology Code: ELEC005Z

Supervisor and instructor: Prof. Paavo Alku

This thesis deals with vocal-effort-focused speaking style conversion (SSC). Specifically, we studied two topics on conversion of normal speech to high vocal effort. The first topic involves the conversion of normal speech to shouted speech. We employed this conversion in a speaker recognition system with vocal effort mismatch between test and enrollment utterances (shouted speech vs. normal speech). The mismatch causes a degradation of the system's speaker identification performance. As solution, we proposed a SSC system that included a novel spectral mapping, used along a statistical mapping technique, to transform the mel-frequency spectral energies of normal speech enrollment utterances towards their counterparts in shouted speech. We evaluated the proposed solution by comparing speaker identification rates for a state-of-the-art i-vector-based speaker recognition system, with and without applying SSC to the enrollment utterances. Our results showed that applying the proposed SSC pre-processing to the enrollment data improves considerably the speaker identification rates.

The second topic involves a normal-to-Lombard speech conversion. We proposed a vocoder-based parametric SSC system to perform the conversion. This system first extracts speech features using the vocoder. Next, a mapping technique, robust to data scarcity, maps the features. Finally, the vocoder synthesizes the mapped features into speech. We used two vocoders in the conversion system, for comparison: a glottal vocoder and the widely used STRAIGHT. We assessed the converted speech from the two vocoder cases with two subjective listening tests that measured similarity to Lombard speech and naturalness. The similarity subjective test showed that, for both vocoder cases, our proposed SSC system was able to convert normal speech to Lombard speech. The naturalness subjective test showed that the converted samples using the glottal vocoder were clearly more natural than those obtained with STRAIGHT.

Keywords: speaking style conversion, high vocal effort, Lombard speech, shouted speech

# Preface

Firstly, I would like to thank my supervisor, Professor Paavo Alku, for giving me the opportunity to work on a relevant field on speech technology, and to learn from his wide knowledge and experience. I would also like to thank Rahim Saeidi, Okko Räsänen, Shreyas Seshadri and Lauri Juvela for their contributions to the published works included in this thesis. In addition, I would like to express my gratitude to Ulpu Remes for her valuable comments, which improved this thesis greatly. I am also grateful for the nice environment created by the speech research groups at Aalto ELEC, first when working at Valotalo, and later in the Health Technology House. It has been specially a pleasure to share office space for many years with Katri Leino. I would also like to thank the examiner of this thesis, Dr. Ville Hautamäki, for his valuable comments and feedback.

Finally, I would like to thank my family, my boyfriend and my friends for their constant support. A special thanks goes to my parents for their support during my academic years, and I also thank my mother for always instilling in me a positive attitude.

Espoo, 30.06.2020

Ana Ramírez López

# Contents

# List of abbreviations

| | |
|---|---|
| **ABE** | aperiodicity band energy |
| **AME** | attenuated main excitation |
| **APLP** | adaptive pre-emphasis linear prediction |
| **ASR** | automatic speech recognition |
| **BGMM** | Bayesian Gaussian mixture model |
| **cycleGAN** | cycle-consistent generative adversarial network |
| **DFT** | discrete Fourier transform |
| **DNN** | deep neural network |
| **DTW** | dynamic time warping |
| **EM** | expectation-maximization |
| **FFT** | fast Fourier transform |
| **GD** | gender-dependent |
| **GIF** | glottal inverse filtering |
| **GMM** | Gaussian mixture model |
| **HMM** | Hidden Markov model |
| **HNM** | harmonics plus noise model |
| **HNR** | harmonic-to-noise ratio |
| **IAIF** | iterative adaptive inverse filtering |
| **INCA** | Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment method |
| **JDE** | joint density estimation |
| **KL** | Kullback-Leibler |
| **LDA** | linear discriminant analysis |
| **LP** | linear prediction |
| **LSF** | line spectral frequency |

| | |
|---|---|
| **LTI** | linear time invariant |
| **MFBE** | Mel-scale filter bank energy |
| **MFCC** | Mel-frequency cepstral coefficient |
| **MGC** | Mel-generalized cepstrum |
| **MI** | mutual information |
| **ML** | machine learning |
| **MLE** | maximum likelihood estimation |
| **MMSE** | minimum mean square estimate |
| **MSE** | mean square error |
| **NNLS** | non-negative least square |
| **OLA** | overlap-add |
| **PLDA** | probabilistic linear discriminant analysis |
| **PML** | pulse model in log-domain |
| **PSM** | perceptual spectral matching |
| **PSOLA** | pitch-synchronous overlap-add |
| **QCP** | quasi-closed phase |
| **RMS** | root-mean-square |
| **SD** | speaker-dependent |
| **SII** | speech intelligibility index |
| **SIIB** | speech intelligibility in bits |
| **SNR** | signal-to-noise ratio |
| **SPL** | sound pressure level |
| **SPSS** | statistical parametric speech synthesis |
| **SSC** | speaking style conversion |
| **TTS** | text-to-speech |
| **UBM** | universal background model |
| **VC** | voice conversion |
| **WER** | word error rate |
| **WLP** | weighted linear prediction |

# List of symbols

## List of Latin symbols

| | |
|---|---|
| $A_m(z)$ | transfer function of inverse filter of $H_m(z)$ |
| $A_{VT1}(z)$ | transfer function of inverse filter of $H_{VT1}(z)$ |
| $\boldsymbol{D}$ | diagonal matrix with elements of $\lvert H_m(\tilde{\Omega})\rvert^2$ in its diagonal |
| $\boldsymbol{E}$ | expansion matrix, to expand $\hat{\boldsymbol{r}}$ to full-length, $N_{FFT}$-point power spectrum |
| $\boldsymbol{\varepsilon}_{target}$ | Mel-scale filter bank energy (MFBE) vector of target speech |
| $f_0$ | fundamental frequency |
| $G(z)$ | $z$-transform of glottal excitation airflow |
| $H_m(z)$ | transfer function of all-pole mapping filter |
| $\lvert H_m(\tilde{\Omega})\rvert^2$ | matching filter power spectrum |
| $h_m(k)$ | impulse response of $H_m(z)$ |
| $H_{VT1}(z)$ | transfer function of 1st-order all-pole filter of vocal tract model |
| $i$ | index for filters $\{\boldsymbol{t}_i\}$ in filter bank $\boldsymbol{T}$ |
| $R_i^c$ | central region of filter $\boldsymbol{t}_i$ |
| $R_i^l$ | lower region of filter $\boldsymbol{t}_i$ |
| $R_i^u$ | upper region of filter $\boldsymbol{t}_i$ |
| $k$ | speech time series index |
| $L(z)$ | transfer function of lip radiation effect |
| $M$ | number of filters in filter bank $\boldsymbol{T}$ |
| $N_{FFT}$ | number of points (bins) in fast Fourier transform (FFT) used |
| $p$ | filter order of $H_m(z)$ |
| $\boldsymbol{r}$ | elementary power spectrum, which is a vector holding the values of segments from piecewise constant power spectrum $\lvert H_m(\tilde{\Omega})\rvert^2$ |
| $r_i$ | $i$th segment of elementary power spectrum $\boldsymbol{r}$ |

| | |
|---|---|
| $\hat{\boldsymbol{r}}$ | estimate of $\boldsymbol{r}$ |
| $\hat{\hat{\boldsymbol{r}}}$ | full-length, $N_{FFT}$-point power-spectrum of vector $\hat{\boldsymbol{r}}$ |
| $S(z)$ | $z$-transform of a speech frame |
| $\lvert S_{source}(\tilde{\Omega})\rvert^2$ | Mel-warped power spectrum of source speech |
| $\lvert S_{target}(\tilde{\Omega})\rvert^2$ | Mel-warped power spectrum of target speech |
| $s_{source}(k)$ | source speech, that is, speech uttered in source speaking style |
| $s_{target}(k)$ | target speech, that is, speech uttered in target speaking style |
| $\boldsymbol{T}$ | uniform-scale triangular filter bank |
| $\boldsymbol{t}_i$ | $i$th filter in filter bank $\boldsymbol{T}$ |
| $V(z)$ | transfer function of vocal tract |
| | |
| $Dir(\cdot)$ | Dirichlet probability distribution |
| $F$ | dimension of feature vectors in $\boldsymbol{X}$, and in $\boldsymbol{Y}$ |
| $j$ | mixture component index |
| $J$ | number of mixture components |
| $\boldsymbol{L}_j$ | precision matrix of $j$th Student t's distribution component |
| $\boldsymbol{m}_0$ | mean vector of the prior distribution (Gaussian-Wishart) over $\boldsymbol{\mu}_j$ |
| $\boldsymbol{m}_j$ | mean vector of the variational posterior distribution (Gaussian-Wishart) over $\boldsymbol{\mu}_j$, and also mean vector of $j$th Student t's mixture component |
| $\mathcal{N}(\cdot,\cdot)$ | Gaussian probability distribution |
| $n$ | observation index |
| $N$ | number of observations |
| $p(\cdot)$ | probability distribution |
| $q(\cdot)$ | variational (approximation) probability distribution |
| $St(\cdot,\cdot,\cdot)$ | Student t's probability distribution |
| $\mathcal{W}(\cdot,\cdot)$ | Wishart probability distribution |
| $\boldsymbol{W}_0$ | scale matrix of the prior distribution (Gaussian-Wishart) over $\boldsymbol{\Sigma}_j$ |

| | |
|---|---|
| $\boldsymbol{W}_j$ | scale matrix of the variational posterior distribution (Gaussian-Wishart) over $\boldsymbol{\Sigma}_j$ |
| $\boldsymbol{w}$ | weight vector of the $J$ Gaussian mixture components. |
| $w_j$ | weight of $j$th Gaussian mixture component |
| $\boldsymbol{X}$ | set of observed feature vectors, from source and target speech of one frame (full training data) |
| $\boldsymbol{x}^{(s)}$ | observed feature vector of a source speech frame (training data) |
| $\boldsymbol{x}^{(t)}$ | observed feature vector of a target speech frame (training data) |
| $\boldsymbol{Y}$ | set of observed and unobserved feature vectors, from source (observed) and target (unobserved) speech of one frame |
| $\boldsymbol{y}^{(s)}$ | observed feature vector of a source speech frame |
| $\boldsymbol{y}^{(t)}$ | unobserved feature vector of a target speech frame |
| $\boldsymbol{Z}$ | set of all latent variables ($\boldsymbol{z}$) and parameters of the Bayesian Gaussian mixture model (BGMM) |
| $\boldsymbol{z}_n$ | latent variable, 1-of-$J$ binary vector associated to $n$th data point |
| $z_{nj}$ | latent variable vector element associated to $n$th data point. |

## List of Greek symbols

| | |
|---|---|
| $\Delta$ | first-order delta coefficients |
| $\Delta\Delta$ | second-order delta coefficients |
| $\zeta$ | Mel-warping coefficient |
| $\tilde{\Omega}$ | index of Mel-warped FFT bins |
| $\boldsymbol{\alpha}_0$ | vector of pseudo observation counts of the prior distribution (Dirichlet) over $\boldsymbol{w}$ |
| $\boldsymbol{\alpha}$ | vector of pseudo observation counts of the variational posterior distribution (Dirichlet) over $\boldsymbol{w}$, and weight vector of the $J$ Student t's mixture components. |

| | |
|---|---|
| $\alpha_j$ | weight of the $j$th Student t's mixture component |
| $\beta_0$ | scale of the prior distribution (Gaussian-Wishart) over $\boldsymbol{\mu}_j$ |
| $\beta_j$ | scale of the variational posterior distribution (Gaussian-Wishart) over $\boldsymbol{\mu}_j$ |
| $\boldsymbol{\theta}$ | parameter set of a BGMM |
| $\boldsymbol{\lambda}$ | parameter set of a Gaussian mixture model (GMM) |
| $\boldsymbol{\Lambda}_j$ | precision matrix of $j$th Gaussian component |
| $\boldsymbol{\mu}_j$ | mean vector of $j$th Gaussian component |
| $\nu_0$ | degrees of freedom of the prior distribution (Gaussian-Wishart) over $\boldsymbol{\Sigma}_j$ |
| $\nu_j$ | degrees of freedom of the variational posterior distribution (Gaussian-Wishart) over $\boldsymbol{\Sigma}_j$ |
| $\boldsymbol{\Sigma}_j$ | covariance matrix of $j$th Gaussian component |

**Comments on notation:**
The symbols on the thesis marked with a hat are an estimate of the given variable. For example, $\hat{\boldsymbol{y}}^{(t)}$ is an estimate of $\boldsymbol{y}^{(t)}$.

# 1 Introduction

Humans can vary their speaking style, and indeed they change it constantly in their daily interactions. Speaking style varies depending on many factors, such as the context of the situation, the state of the speaker, or the personal and social relationship of the speaker with the other interlocutors of the conversation. Thus, given the ubiquity of speaking style in natural speech, it is important that speech technology adapts to it with the objective of obtaining more realistic results. One way to achieve this is by including speaking style conversion (SSC) as a part of the steps performed in the current speech technology. SSC performs an acoustic-to-acoustic conversion of the original speaking style of a speech utterance (denoted henceforth as the *source speaking style*) to another speaking style of our choice (denoted henceforth as the *target speaking style*). Then, for example, whispered speech could be converted to shouted speech, or normal (neutral) speech could be converted to sad speech (that is, speech uttered in sad emotion).

If we pay attention to the different aspects in which someone's speaking style may change, we can notice that the style can vary for example in terms of emotion and/or of vocal effort. When performing SSC that is focused on emotion, often denoted plainly as *emotion conversion* (e.g., [1, 2, 3, 4]), the focus in conversion is mainly in paralinguistic attributes of speech, such as prosody and intonation. In case of SSC for vocal effort, speech attributes such as energy/intensity, loudness and pitch of the signal become more important for an optimal conversion. Nevertheless, these two aspects of conversion (emotion and vocal effort) are not entirely separate, and become sometimes completely intermingled; for example, in case of angry speech [5]. Given all the aforementioned changes in speech attributes for different speaking styles, the main challenge of SSC tasks is to achieve speech conversion by transforming (some of) those attributes while at the same time retaining the voice (that is, the speaker identity) and the linguistic content of the utterance. In addition, it is essential that the SSC system does not sacrifice speech quality to achieve converted samples that show a clear target speaking style. That is, rather than having a compromise between speech quality and degree of conversion, it is desirable to have a SSC system that achieves to have both.

SSC applications include those where the end user is a human listener and those where the end user is a machine learning (ML) system. In the case of vocal-effort-focused SSC, applications intended for human listeners include making the converted speech signal more intelligible. For example, soft speech (such as whispered speech) could be converted to normal speech to make it more understandable, or normal speech could be converted to Lombard speech [6] in order to make it more intelligible when listened to in noisy situations. In addition, normal speech could be converted to so-called clear speech [7, 8], which facilitates comprehension for the listener. These applications are most beneficial for people with hearing impairments or for people who have difficulties in understanding speech produced using normal speaking style. Another potential application of vocal-effort-focused SSC (and SSC in general) for human listeners, could be the customization of speech according to the preferences of the end user. For example, text-to-speech (TTS) concatenative synthesis systems'

output speech could become more flexible and personalizable in accordance to the user needs, without the need of a bigger data set. This could be achieved by applying SSC to the synthesized samples [9, 10]. On the other hand, some SSC solutions could be employed in a parametric speech synthesis system, by deploying rules based on SSC to perform speaking style modifications [11].

Vocal-effort-focused SSC can prove to be useful also for speech computer-based tasks, as we will see next. Humans have the innate skill of being able to cope with variation in speech (for example, changes in accent, vocal effort, or voice mimicry) during everyday tasks, such as speech or speaker recognition. In contrast, variations in speech recordings pose a challenge for computer-based systems. Typically, studies comparing the performance of humans and machines at speech-related tasks have shown that humans usually outperform machines [12, 13, 14]. Nevertheless, the advances in speech technology over the years have decreased the performance gap between machines and humans; in some cases, machines managed to equal or even slightly surpass humans. The latter has been shown for example in some studies on speaker recognition or verification for voice mimicry or disguise [15, 16, 17] and also in some studies on speech recognition [18]. In case of vocal effort variations, a common case of mismatch occurs for speaker recognition or verification tasks in forensic cases: the system is usually trained on normal speech, while the speech samples under evaluation are sometimes uttered by a speaker that is in an agitated or stressed state. Such mismatch harms the performance of the system [19, 20]. By decreasing the mismatch, the recognition performance could improve [21, 22, 23]. Thus, vocal-effort-focused SSC can be applied in such cases. We should note that SSC applications oriented to ML systems differ from those oriented to human listeners in that the converted speech samples do not require to retain the subjective speech quality of the original signal.

SSC methods use mainly two approaches: a *vocoder-based parametric approach*, and a *direct transformation approach*. The vocoder-based parametric approach employs a vocoder to extract features from a speech signal, a subset of those features are then modified, and finally the vocoder, having as input all features (modified and unmodified), synthesize the converted speech signal. In contrast, the direct transformation method involves converting directly the source speech signal to target speech, by applying operations (such as filtering) directly onto the speech signal, or onto the speech signal once transformed to another domain (e.g. spectral domain). The transformation of the features can be automatic, using ML-based mapping techniques. Some mapping methods employed for SSC require having parallel data for fitting their models. Obtaining parallel data for speaking styles is quite costly in general and thus this kind of databases are scarce. Therefore, some mapping techniques have been used in SSC specifically to cope with the problem of needing parallel data for mapping.

SSC has relation with other areas in the speech technology field, such as voice conversion (VC) [24], statistical parametric speech synthesis (SPSS) [25] and speech enhancement in speech transmission [26]. Nevertheless, SSC can be understood as a research area on its own due to its differences with the other aforementioned areas. For example, there is no linguistic-to-acoustic conversion as in SPSS. On the

other hand, SSC is not constricted by strict latency constraints which are present in enhancement applications in speech transmission technology.

## 1.1 Thesis scope

This thesis focuses on speaking styles that differ in terms of vocal effort. Specifically, the focus is on conversion from normal speaking style to a speaking style uttered in high vocal effort. Based on this aim, we studied two different topics on SSC in this thesis: the first topic dealt with normal-to-shouted speech conversion and the second topic dealt with normal-to-Lombard conversion. In this thesis, we present experimental work from these two topics based on the following peer-reviewed articles:

- [22] A. Ramírez López, R. Saeidi, L. Juvela, and P. Alku, "Normal-to-shouted speech spectral mapping for speaker recognition under vocal effort mismatch." in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4940–4944.
- [27] A. Ramírez 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 *Interspeech*, 2017, pp. 1363–1367.

The SSC system that we proposed in the first topic used a direct transformation approach, that we denoted as perceptual spectral matching (PSM)-Gaussian mixture model (GMM) (see Sections 5 and 6). This system had direct application in a speaker recognition framework in which there is mismatch of vocal effort between the enrollment and test utterances. In the work we have done on this first topic, the main research question under study was to test how effective SSC would be when used together with a computer-based speech system (in this case, a speaker recognition system), rather than having as end-user a human listener. In order to answer this research question, we evaluated the speaker recognition system performance in two different cases: with and without employing SSC. A side research question consisted of studying if the SSC system was able to convert speech without degrading speaker identity. The SSC system that we proposed for this topic has the novelty of being, to the best of our knowledge, the only SSC work proposed for direct application in speaker recognition with vocal effort mismatch.

The SSC method that we proposed for the second topic, was a vocoder-based parametric SSC system (see Sections 7 and 8). We evaluated this method in terms of speech quality using the converted speech samples in subjective listening tests. With this evaluation, we mainly focused on studying if it is possible to achieve adequate conversion of speech (in the specific case of normal-to-high-vocal-effort conversion) without producing degradation in quality. A secondary research question in this topic was to evaluate if the newer version of a glottal vocoder proposed in [28] that we employed would provide better speech quality in the converted samples than the widely used STRAIGHT vocoder. In addition, a minor research question was to test if the mapping system we selected (Bayesian Gaussian mixture models (BGMMs))

would function adequately and cope with the scarcity of the training data. Finally, we should note that the work we have performed on this topic has the novelty of using BGMM-based mapping for the first time in a SSC study.

## 1.2 Thesis structure

The present thesis consists of a theoretical background and experimental work in SSC of normal speech to high vocal effort speech. We present the theoretical background in Sections 2-4. Section 2 covers details about the speech production process and developed theory for modelling it, which is a foundation to the vocoders employed in the second topic of this thesis. We present also details of the vocoders used. In addition, given that we are focusing in this thesis on conversion of normal speech to high vocal effort speech, this section also describes differences in speech attributes between these styles of speech. Section 3 introduces a brief overview of the mapping techniques employed for SSC, and it goes more in depth onto the mapping techniques that we employed in the experimental works included in this thesis. Finally, Section 4 presents a review of high-vocal-effort-focused SSC.

We present the experimental work for each topic, along with theory of the SSC methods employed in each topic, in Sections 5-8. We organized the sections as follows. First, Section 5 presents the theory of the first topic: we proposed a direct-transformation-based SSC system for SSC of normal speech to shouted speech, with a speaker recognition application under vocal effort mismatch. This study corresponds to work published in [22]. We then present the experimental work on this topic in Section 6. Second, Section 7 introduces the theory of the second topic: a vocoder-based parametric SSC system we proposed, to transform normal speech to Lombard speech. The work from this second topic has been previously published in [27]. We present the experimental work of the second topic in Section 8. Finally, Section 9 presents overall discussion and conclusions from this thesis.

# 2 Speech production and its modeling

## 2.1 The speech production mechanism

If we think of the mechanism of human speech production from a signal point of view, we can consider speech production as a two-step process. In the first step, we generate airflow by exhaling air from our lungs through the trachea, and then the airflow passes through the vocal folds at the larynx. If our vocal folds are open when the airflow passes through, an excitation signal is created corresponding to unvoiced speech. In the case of voiced speech, the vocal folds are tense, vibrate and collide, such that the airflow signal is modulated. As a result, an excitation signal, in the form of quasi-periodic pulses, is generated. This voiced excitation signal is frequently referred to as *glottal flow*[1] or voice source. The rate at which the vocal folds vibrate indicate the value of $f_0$ (the *fundamental frequency*), which is the pulse frequency of the glottal flow signal. Glottal flow signal thus generates the harmonic spectral structure of speech given by $f_0$ and its harmonics. In contrast, the unvoiced excitation lacks the harmonic structure, because the vocal folds do not vibrate periodically.

In the second step, the excitation signal passes through our vocal tract and is radiated via our mouth and nostrils. The vocal tract consists of the following articulators: pharynx, oral and nasal cavities. The vocal tract filters the excitation signal, since the vocal tract articulators act as cavity resonators and create resonances (also called *formants*) that shape the excitation signal in the spectral domain. We can modify at will the resonances by changing the position or shape of our vocal tract articulators. Thus we humans are capable of creating speech signals of different formant values which is very important for recognition of phonemes. The vocal tract dimensions and shape vary across gender, age, and even across individuals [29, 30]. In consequence, even if a group of people would utter speech sounds representing the same phoneme, the produced signals would show differences in spectral content and the formants of the signals would also not be exactly same. In the case of unvoiced speech, the airflow is constricted partially (or totally for a short instant for the so-called plosives) at some point of the vocal tract. As result, unvoiced speech shows a noise-like waveform (or impulse-like for plosives). Figure 1 presents a simplified sketch of the human speech production mechanism.

## 2.2 Source–filter modeling

The two-step notion of speech production, covered in Section 2.1, is the foundation of source–filter theory [29], widely used in speech modeling. The speech production mechanism is modeled as an excitation signal (the source) that excites the vocal tract (the filter) and this one in turn shapes the spectrum of the signal. Figure 2 presents a block diagram of speech production modeling with source–filter theory.

[1] In acoustics, the correct term for the the voiced excitation signal is the glottal volume velocity waveform, but it is commonly just referred as the glottal flow. The 'glottal' naming references the V-shaped opening between the vocal folds, denoted as glottis.

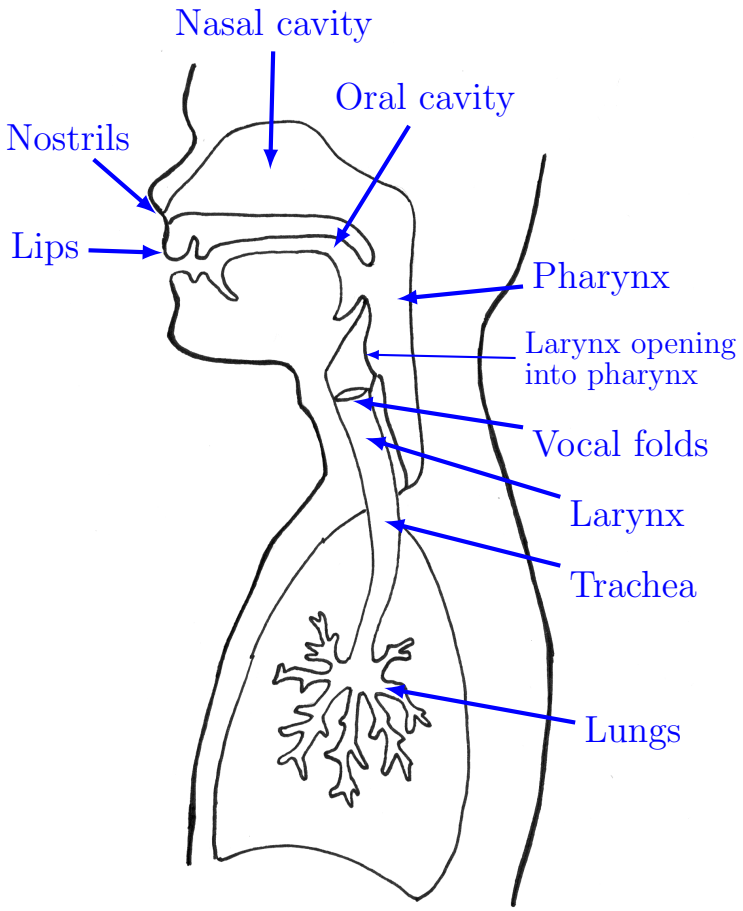


Figure 1: Speech production mechanism.

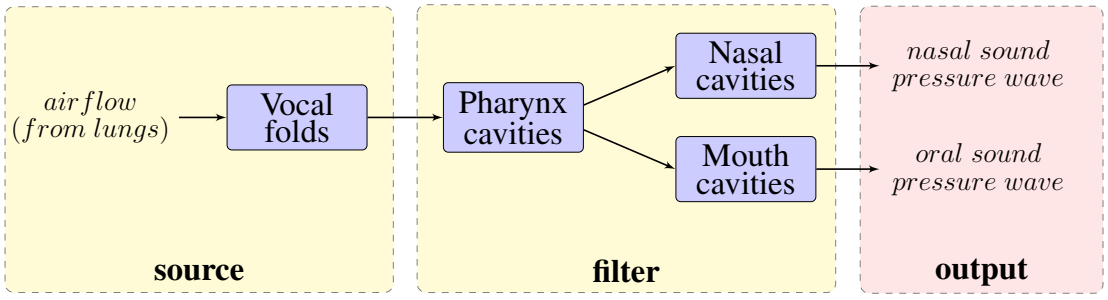


Figure 2: Schema of source–filter model of speech production, adapted from [29, 31].

In the case of voiced speech, the excitation signal is the glottal flow, which is the result of vocal folds vibrating when airflow passes through them. Thus, the glottal flow spectrum presents harmonics and the magnitude spectrum decreases with the frequency, at a rate of 12dB per octave [32]. In the case of unvoiced speech, we can assume the excitation signal to be white noise, since for unvoiced speech the airflow is constricted somewhere in the vocal tract which results in a noise-like signal. The

vocal tract (seen as an acoustic tube) modulates the source signal and generates formants that amplify the spectrum magnitude of the source signal at the formant's frequency and its surrounding[2]. After passing through the vocal tract, the signal radiates through the mouth and nose as a sound pressure wave and we are able to hear it. When the conversion from airflow to pressure wave occurs, a *lip radiation effect* is observed, resulting from the change in acoustic impedance between the lips and the surrounding air. We can approximate the lip radiation effect by the first derivative of the airflow, so it behaves like a high-pass filter and the spectrum magnitude increases at a rate of 6dB per octave [32]. Thus, in a digital system we can approximate the lip radiation effect to a first-order differentiator, which in the $z$ domain is represented with the following transfer function:

$$L(z) = 1 - \gamma z^{-1}, \tag{1}$$

where $\gamma$ is a constant value. Usually $\gamma \leq 1$, to ensure stability when the lip radiation effect needs to be cancelled by inversion.

The source–filter model assumes that the two components that conform speech, source and filter, are independent of each other and thus they can be computed independently. In addition, the model assumes the filter to be linear time invariant (LTI). Given these assumptions, speech can be represented using the source–filter model, in the $z$ domain, as [29, 33]:

$$S(z) = G(z)V(z)L(z), \tag{2}$$

where $G(z)$ is z-transform of the glottal source signal, $V(z)$ is the transfer function of the vocal tract, and $L(z)$ is the transfer function of the lip radiation (Eq. 1). The lip radiation effect is frequently combined with the glottal source component, and then speech can be expressed as:

$$S(z) = G'(z)V(z), \tag{3}$$

where

$$G'(z) = G(z)L(z). \tag{4}$$

The source–filter model is a simplification of the actual speech production mechanism, and therefore has some drawbacks. The main flaw comes from the assumption of independence between the source and filter elements of the model, while in reality there is interaction effects between these two [34, 35]. Another flaw involves the assumption of having a LTI filter: while the model is accurate enough for sounds that vary slowly (like voiced speech), there are other sounds (for example, plosive consonants like /p/ or /k/) which are more rapidly changing. In such cases, the model fails at representing speech accurately. In addition, vocal tract is commonly represented with an all-pole filter (which is frequently computed using linear prediction (LP) [36]). Thus, sounds with anti-resonances (like nasal sounds), which require having zeros in the filter, will be poorly modeled. Nevertheless, we can solve

[2]In the case of nasal sounds, anti-resonances can also be created, which attenuate the spectrum magnitude.

this issue by adding extra poles to the filter [37]. We should also note that given the coupling effects between filter and source that the model neglects, the all-pole filter is most likely modeling not only the vocal tract but also some contributions of the source and lip radiation effect. The aforementioned flaws limit the accuracy of the source–filter model. The coupling effects between source and filter may affect specially the performance of speech applications focusing on generating speech or modifying it, since for those applications the naturalness in speech is a key matter. In other applications, such as speech coding, source–filter modeling proves to be good enough.

## 2.3 Source–filter vocoders

SPSS tasks have at their core a vocoding system: 1) A vocoder is used during the training stage, to extract features that represent the speech signal. The features are then used to train statistical generative models. These models are indexed with a linguistic specification, which gives textual context information, and it is stored for later retrieval. 2) A vocoder is also used in the synthesis stage. In this phase, linguistic specifications extracted from text are used as input to retrieve the trained statistical models, which in turn estimate the speech features. Finally, these features are used as input to the vocoder, to reconstruct the speech signal [38, 39]. As mentioned in Section 1, while speech technology applications such as SPSS, VC and SSC have different goals, there is some relation between them. In VC and SSC a vocoder is often employed also as part of the system.

Most commonly, vocoders are based on the source–filter model (e.g. STRAIGHT [40, 41], WORLD [42], GlottHMM [43], GlottDNN [44], pulse model in log-domain (PML) [45] and GSS [46]). There are also vocoders employing other models: for example, using the harmonic (or sinusoidal) model, which represents speech as a sum of sinusoids [47] (e.g. Ahocoder [48] and HMPD [49]) or the dynamic sinusoidal model, in which a time-varying term is added to the standard sinusoidal model for amplitude refinement [50] (e.g. PDM [51, 50]).

In the experimental work from the second topic included in this thesis, we employed two source–filter vocoders: a glottal vocoder and STRAIGHT. Thus, these two vocoders are explained in detail next.

### 2.3.1 The glottal vocoder

Glottal vocoders are based on the source–filter parametric model [29], such that speech can be represented as a convolution of the vocal tract filter and glottal flow excitation (the latter one includes also the lip radiation effect) [29, 52]. For the second topic of this thesis, we employed a glottal vocoder that is a variant implementation of the glottal vocoder introduced in [28]. This vocoder was created at first for SPSS applications [25]. It employs for voiced frames a glottal inverse filtering (GIF) method based on the source–filter model to split the speech signal into a vocal tract filter and glottal flow excitation. GIF methods are able to estimate the glottal flow of the speech signal by cancelling out the effect of the vocal tract and

lip radiation. When the source–filter model is assumed for speech production (see Eqs. 2-3), GIF estimates the glottal flow in a voiced segment as:

$$G(z) = \frac{S(z)}{V(z)L(z)}. \tag{5}$$

where G(z) is the z-transform of the glottal flow, S(z) is the z-transform of the speech signal, V(z) is the transfer function of the vocal tract and L(z) the transfer function of the lip radiation effect. As seen in Eq. 1, we can express lip radiation as a first-order differentiator; thus, the main task is to estimate the vocal tract element accurately. A frequent issue when estimating the vocal tract is harmonic bias: formants are affected by this bias and their estimates tend to shift towards harmonics generated by the voice source. This is specially true for high-pitched signals (such as those uttered by female speakers), which show sparse, high-energy harmonics.

The glottal vocoder that we employed here uses specifically the quasi-closed phase (QCP) GIF method [53], which is based on closed phase analysis. In this approach, the vocal tract spectrum is estimated during the closed phase of the glottal excitation, i.e. when the glottis is closed. At this time, the voice source's influence on the vocal tract's spectrum is minimal. QCP uses weighted linear prediction (WLP) [54] and the attenuated main excitation (AME) weight function that minimizes the biasing effect from the harmonics of the glottal flow signal when estimating the vocal tract's spectrum [53]. In the case of unvoiced segments, the vocoder uses a random noise excitation signal and conventional LP for the vocal tract. In addition, this vocoder uses a deep neural network (DNN) model to generate the glottal flow pulses [55], which are employed in the synthesis step. Lastly, to parametrize speech, the glottal vocoder extracts during analysis the following features: 1) log-energy, 2) harmonic-to-noise ratio (HNR), 3) $f_0$, 4) vocal tract line spectral frequencies (LSFs), denoted here as $LSF_{VT}$, and 5) glottal source LSFs, denoted as $LSF_{glott}$.

### 2.3.2 The STRAIGHT vocoder

STRAIGHT is a known vocoder, often used in SPSS, which also uses as foundation the source–filter model [29]. The STRAIGHT vocoder estimates during analysis a cepstrum-based spectral envelope for the vocal tract using a pitch-adaptive time-frequency smoothing method. This method also aims to minimize the biasing effect generated by the harmonic peaks to the vocal tract spectrum [40, 41]. In addition, STRAIGHT employs a mixed excitation signal during synthesis. This kind of signal involves: 1) a periodic train of pulses, mixed with 2) an aperiodic noise signal, which is added to several frequency bands based on some aperiodicity weights. The mixed excitation signal is used for voiced segments, while a white Gaussian noise excitation signal is used for unvoiced segments. Finally, the features that this vocoder extracts during the analysis stage are: 1) the aperiodicity band energies (ABEs), to represent the aperiodicity spectrum, 2) $f_0$, and 3) the spectral envelope, which is represented using a Mel-generalized cepstrum (MGC).

## 2.4 High vocal effort speech

When the focus is on speech uttered using different vocal efforts, we can consider that the different vocal effort modes define a kind of continuum, from low to high vocal effort. There are many examples of vocal effort modes on the continuum such as whispered speech, soft speech, normal speech, loud speech, Lombard speech, and shouted speech. However, we should note that these examples of vocal effort modes are not completely separate and acoustical speech features belonging to the different modes on the continuum typically overlap. In this thesis, our focus is on speaking styles of high vocal effort, which are compared to the speaking style used in production of speech of normal vocal effort.

For efficient SSC, the system should use speech attributes that most prominently differ between the source speaking style and the target speaking style. In the experimental work of this thesis, the source speaking style was in all cases normal speech, while the target speaking style was Lombard speech in the second topic and shouted speech in the first topic. Next, we present the speech attributes of both target speaking styles (Lombard speech and shouted speech), and compare them to those of the source speaking style (normal speech).

### 2.4.1 Lombard speech

There are certain speech production mechanisms that humans employ to enhance the intelligibility of their speech. For example, humans increase their vocal effort (and in turn they increase their loudness) to be heard more easily. The increase in vocal effort is an involuntary reflex, known as the Lombard effect [6, 56, 57], that humans adopt when they are placed in adverse acoustic environments (i.e. noisy conditions) [58].

The difference in loudness between normal and Lombard speech is the most evident speech attribute that changes between these two styles. However, there are also other changes affecting the speech. One is the difference in duration between normal and Lombard speech: there is an increase in duration for Lombard speech in comparison to its normal speech counterpart [58, 59, 60]. Another speech attribute difference is $f_0$, which also increases for Lombard speech [58, 59, 61]. In addition, the spectral content of the speech signal varies from one style to the other. Specifically, in Lombard speech, the energy of higher frequencies is typically larger compared to normal speech. In other words, the spectrum of Lombard speech shows a flatter tilt compared to normal speech [60, 61]. Within all these attributes, spectral tilt has been shown to influence most for the intelligibility enhancement provided by Lombard speech [61]. Finally, we should note that the type of noise that triggers the Lombard effect [60] and the gender of the speaker [62] affect the quantity and manner in which these speech attributes change.

### 2.4.2 Shouted speech

Humans utter in shouted speaking style when for example a person tries to communicate with another person over a distance [63], or when the speaker is in an agitated

or stressful state [64]. We can encounter the latter situation often in forensic cases [65]. While shouted speech is at the end of the vocal effort continuum, and thus shows typically a higher sound pressure level (SPL) value than Lombard speech, the use of shouted speech is less intelligible compared to Lombard speech, or normal speech. The reason for this is the reduced use of articulation during the production of shouted speech [66, 67].

Apart from the increase in SPL [20], also other speech attributes change for shouted speech in comparison to normal speech. For example, $f_0$ increases in shouted speech [68, 69, 65] as well as the first formant ($f_1$). [70, 69, 65]. Regarding the spectral distribution, the spectral tilt in shouted speech decreases in comparison to normal speech [20, 71, 69]. On the other hand, the duration of utterances produced using shouting increases in comparison to the same utterances uttered in normal speaking style [68, 20]. This increase in duration is due to an increase in word duration, rather than in silence duration (as it occurs in whispered or soft speech) [20].

# 3 Mapping techniques

Several approaches exist for mapping speech features of the source speaking style to speech features of the target speaking style. Some SSC methods employ signal processing methods only in their approach, while other methods employ a combination of signal processing and ML techniques in the conversion. The SSC methods that only employ signal processing can be understood as deterministic approaches, given that in that case we are applying a fixed solution to all the speech input signals. Among high-vocal-effort-focused SSC works, [11, 72, 10, 73] employ only signal processing. Out of these, [10] studies a scenario in which parallel data of the source and target speech is required, since vocal effort modification is based on transferring features of target speech to source speech. While deterministic approaches, like the aforementioned ones, tend to be computationally cheaper, data-driven SSC approaches (that is, approaches that include ML along signal processing) are usually more effective due to their flexibility: the solution proposed is adaptive to the given data, rather than fixed. The adaptation is based on the trained model that the data-driven approach employs. Thus, data-driven mapping techniques require training data of the source and target signals to fit in the model.

In the case of speech applications, some mapping techniques require to have a data set of parallel training speech samples. That is, pairs of speech samples that have the same attributes (such as text, voice, or speaking style), except for the attribute that is being mapped: one sample out of the pair will have the to-be-mapped attribute from source speech, while the other sample will have that attribute from target speech. For SSC, the training data set consists of pairs of speech samples for the source and target that has the same voice identity (speaker) and linguistic content (that is, the same text), but one sample is uttered by the speaker in the source speaking style, and the other samples is uttered in the target speaking style. Other mapping techniques can be trained with non-parallel data. This means that the speech samples from source and target speech do not need to correspond to the same linguistic contents. For this reason, parallel data is also referred to as text-dependent data, while non-parallel data is referred to as text-independent data. Based on the type of data required in training, we can group data-driven mapping techniques into two main categories: parallel mapping and non-parallel mapping.

## 3.1 Data-driven, parallel mapping

Many data-driven mapping approaches for SSC have been adopted from the VC field, in which more research has been conducted. In the case of parallel-based approaches, GMMs have been, for example, used for mapping in SSC, and were firstly proposed in VC [74, 75]. In high-vocal-effort-focused SSC studies, GMMs have been applied in [9, 22]. GMMs were also employed in [76], and were compared against feed-forward DNNs, and BGMMs; the latter ones are an extension of GMMs, and are more robust to scarce training data than GMMs. BGMMs have also been used earlier in [27]. Robust mapping techniques are key in case of SSC, since collecting data in different speaking styles is very costly; this is specially true for parallel data.

In a parallel data set, often the pairs of source and target speech signals do not match in duration. Thus, prior to using the parallel data set in mapping, its source-target speech pairs need to be time aligned. A common approach for alignment at frame-level is dynamic time warping (DTW) [77, 78, 79, 74]. Other approaches are those that perform Hidden Markov model (HMM)-based phonetic aligments [80, 81] or sentence HMM-based alignments [82]. The performance of the alignment task, to pair the source and target frames, also influences the outcome of the conversion tasks. This topic is not covered in the present thesis, but it has been studied for example in [83].

In this thesis, we present two topics that include data-driven SSC systems. These topics involve mapping using GMMs in one case (first topic) [22], and BGMMs in the other (second topic) [27]. Thus, next we describe in detail these two kinds of statistical models.

### 3.1.1 GMM mapping

GMM is a probabilistic model widely used to represent continuous features in speech technology methods such as in automatic speech recognition (ASR) [84] or VC [74]. GMMs are continuous, parametric density functions that consist of a weighted sum of Gaussian distributions (denoted as the *mixture components* of the GMM). When we employ a GMM to model the observed data $\boldsymbol{X}$, each observation ($\boldsymbol{X}_n$, where $n = 1, 2, \ldots, N$, and $N$ is the total number of observations available) we are assuming each observation to come from one of the Gaussian components of the GMM, though it is unknown to us from which specific component. In other words, we are assuming that $\boldsymbol{X}$ presents a hidden cluster-like structure and the GMM's components model the underlying, latent classes in the data (with each of these classes assumed to follow a Gaussian distribution) [85]. Thus, in this model, each observation sample ($\boldsymbol{X}_n$) has associated a latent variable ($\boldsymbol{z}_n$) that indicates from which mixture component the data point originates. Latent variable $\boldsymbol{z}_n = [z_{n1}, z_{n2}, \cdots, z_{nJ}]$ is a binary vector, in which all elements are random binary variables (that is, $z_{nj} \in \{0, 1\}$), and $J$ is the number of Gaussian components of the model. Of all the vector elements in $\boldsymbol{z}_n$, only one is 1, for example $z_{nj'}$=1, while the rest of vector elements in $\boldsymbol{z}_n$ are 0. This indicates that the corresponding data sample $\boldsymbol{X}_n$ belongs to the $j'$th component of the GMM. Thus, the $J$ elements of vector $\boldsymbol{z}_n$ always add up to 1: $\sum_{j=1}^{J} z_{nj} = 1$.

We can express the likelihood of $\boldsymbol{X}$ as:

$$p(\boldsymbol{X}|\boldsymbol{\lambda}) = \sum_{j=1}^{J} p(z_j = 1|\boldsymbol{\lambda}) p(\boldsymbol{X}|z_j = 1, \boldsymbol{\lambda}), \tag{6}$$

where $\boldsymbol{\lambda}$ are the GMM parameters, $p(z_j = 1|\boldsymbol{\lambda})$ is the prior probability of the $j$th mixture component, and $p(\boldsymbol{X}|z_j = 1, \boldsymbol{\lambda})$ is the $j$th component density function. Given that: 1) $p(z_j)$ is often denoted as $w_j$, that is:

$$p(z_j = 1|\boldsymbol{\lambda}) = w_j, \tag{7}$$

and 2) the probability density functions of each GMM component is a Gaussian distribution:

$$p(\boldsymbol{X}|z_j = 1, \boldsymbol{\lambda}) = p(\boldsymbol{X}|\boldsymbol{\lambda}_j) = \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j); \tag{8}$$

then, using Eqs. 7-8, we can express the likelihood from Eq. 6 as:

$$p(\boldsymbol{X}|\boldsymbol{\lambda}) = \sum_{j=1}^{J} w_j \mathcal{N}(\boldsymbol{X}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad \sum_{j=1}^{J} w_j = 1 \tag{9}$$

where $w_j = p(j|\boldsymbol{\lambda})$ is the mixture weight of the $j$th component, (as mentioned, the prior probability of $j$th mixture component); and $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the mean vector and covariance matrix, respectively, of the $j$th component. In addition, $J$ is the number of mixture components. Thus, the parameters of the model to estimate are $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{w}\}$, where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_j\}$, $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_j\}$, and $\boldsymbol{w} = \{w_j\}$, for $j = 1, 2, \ldots, J$. The estimation of the GMM parameters is typically computed by using maximum likelihood estimation (MLE). Figure 3 shows a graphical model of a GMM.

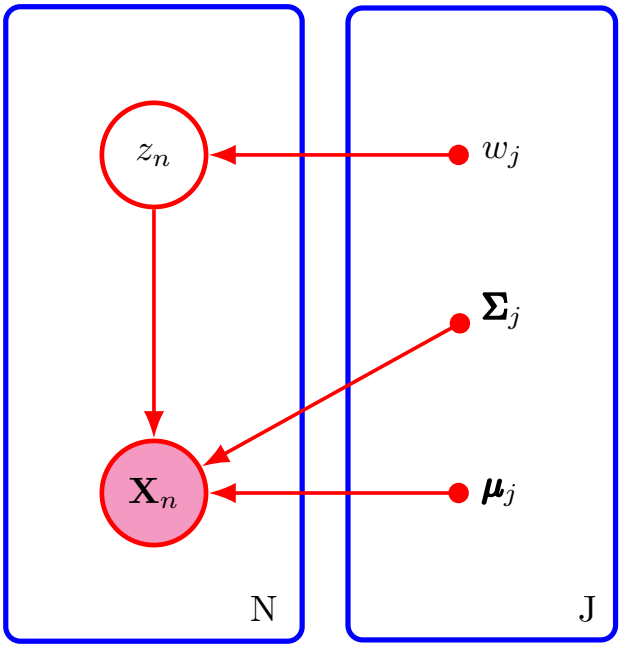


Figure 3: Graphical model of a GMM, adapted from [86]. Nodes represent variables; the node marked in pink corresponds to a observed variable. In addition, red notches indicate point estimates of parameters (that is, fixed values), and arrows represent conditional dependencies. Finally, the plates indicate repetition over the respective index. Thus, on one hand there is a set of $N$ i.i.d. observed data points $\{\boldsymbol{X}_n\}$, with corresponding latent points $\{z_n\}$, with $n = 1, 2, \ldots, N$; and on the other hand there are the parameters $\{w_j\}$, $\{\boldsymbol{\mu}_j\}$, and $\{\boldsymbol{\Sigma}_j\}$, of mixture components $j = 1, 2, \ldots, J$.

GMMs have been used firstly in the context of VC (e.g. [75]), and later in SSC (e.g. [22]), because of the cabalities of GMMs to model the dependencies between

the feature vectors of a source speech frame ($\boldsymbol{x}^{(s)}$) and the feature vectors of a target speech frame ($\boldsymbol{x}^{(t)}$). The joint density estimation (JDE) approach [75], often used with GMMs, implies that concatenated feature vectors $\boldsymbol{X} = [(\boldsymbol{x}^{(s)})^\intercal, (\boldsymbol{x}^{(t)})^\intercal]^\intercal$ are employed for training the GMM (that is, estimating the GMM parameters $\boldsymbol{\lambda}$) that models the joint density function of source and target features, $p(\boldsymbol{X}|\boldsymbol{\lambda})$. As we can appreciate in Eq. 9, this density is the likelihood of $\boldsymbol{X}$, and we can use it afterwards for mapping feature vectors of source speech to those of target speech. In this case, we can rewrite Eq. 9 as follows:

$$p(\boldsymbol{X}|\boldsymbol{\lambda}) = \sum_{j=1}^{J} w_j \mathcal{N}\left( \begin{bmatrix} \mathbf{x}^{(s)} \\ \mathbf{x}^{(t)} \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu}_j^{(s)} \\ \boldsymbol{\mu}_j^{(t)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_j^{(s,s)} & \boldsymbol{\Sigma}_j^{(s,t)} \\ \boldsymbol{\Sigma}_j^{(t,s)} & \boldsymbol{\Sigma}_j^{(t,t)} \end{bmatrix} \right), \quad \sum_{j=1}^{J} w_j = 1 \tag{10}$$

where $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_j^{(s)} \\ \boldsymbol{\mu}_j^{(t)} \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_j^{(s,s)} & \boldsymbol{\Sigma}_j^{(s,t)} \\ \boldsymbol{\Sigma}_j^{(t,s)} & \boldsymbol{\Sigma}_j^{(t,t)} \end{bmatrix}$.

We can compute maximum-likelihood estimates of the model parameters $\boldsymbol{\lambda}$ using the expectation-maximization (EM) algorithm [87]. This is an iterative algorithm, which (after an initial estimate of the model parameters) alternates between two steps: 1) Expectation step: computing the expected values of $\boldsymbol{z}_n$, given the current estimate of model parameters $\boldsymbol{\lambda}$. The expectations of $\{\boldsymbol{z}_n\}$ are denoted as responsibilities, i.e. $r_n = \mathbb{E}(\boldsymbol{z}_n)$. 2) Maximization step: updating the model parameters $\boldsymbol{\lambda}$ with their best maximum-likelihood-based estimates, while keeping fixed the values of the current responsibilities. The iteration between these two steps continues until the algorithm reaches convergence [87]. Some of the disadvantages of EM is that it is very sensitive to the initialization of the parameters, and it requires to determine the number of GMM components (clusters) beforehand. We can interpret EM as a soft version of the $k$-means algorithm. Apart from the disadvantages coming from the EM algorithm, using GMMs also has some downsides. The main disadvantage of GMMs is that given that the parameters are fixed, there is not indication of the uncertainty of their estimated values.

After we have trained the GMM with the EM algorithm, we can predict the target feature $\hat{\mathbf{y}}^{(t)}$ as the mean square error (MSE) estimate of $\mathbf{y}^{(t)}$ given data $\mathbf{y}^{(s)}$:

$$\hat{\boldsymbol{y}}^{(t)} = \sum_{j=1}^{J} p(j|\boldsymbol{y}^{(s)}, \boldsymbol{\lambda}) \left[ \boldsymbol{\mu}_j^{(t)} + \boldsymbol{B}_j (\boldsymbol{y}^{(s)} - \boldsymbol{\mu}_j^{(s)}) \right], \tag{11}$$

where $\boldsymbol{\mu}_j^{(t)}$ and $\boldsymbol{\mu}_j^{(s)}$ are the mean vectors of the $j$th component for $\mathbf{x}^{(t)}$ and $\mathbf{x}^{(s)}$, respectively. We can compute the posterior component probabilities $p(j|\boldsymbol{y}^{(s)}, \boldsymbol{\lambda})$ from prior component probabilities $w_j$ and likelihoods $\mathcal{N}(\boldsymbol{y}^{(s)}|\boldsymbol{\mu}_j^{(s)}, \boldsymbol{\Sigma}_j^{(s,s)})$ as

$$p(j|\boldsymbol{y}^{(s)}, \boldsymbol{\lambda}) = \frac{w_j \mathcal{N}(\boldsymbol{y}^{(s)}|\boldsymbol{\mu}_j^{(s)}, \boldsymbol{\Sigma}_j^{(s,s)})}{\sum_{j'=1}^{J} w_{j'} \mathcal{N}(\boldsymbol{y}^{(s)}|\boldsymbol{\mu}_{j'}^{(s)}, \boldsymbol{\Sigma}_{j'}^{(s,s)})}, \tag{12}$$

and we obtain the linear transformations $\boldsymbol{B}_j$ as

$$\boldsymbol{B}_j = (\boldsymbol{\Sigma}_j^{(t,s)})^{-1} \boldsymbol{\Sigma}_j^{(s,s)}, \tag{13}$$

where $\mathbf{\Sigma}_j^{(t,s)}$ and $\mathbf{\Sigma}_j^{(s,s)}$ are correlation matrices concerning $\mathbf{x}^{(t)}$ and $\mathbf{x}^{(s)}$.

### 3.1.2 BGMM mapping

BGMMs are an extension of GMMs that give better generalization performance, since BGMMs adapt their complexity depending on the data at hand. That is, rather than having a pre-fixed number of components as in GMMs, the number of components are stochastically selected as a function of the data structure. Another advantage of using a Bayesian framework is that it is less sensitive to parameter initialization.

In conventional GMMs, the parameters or the model are deterministic, and we employ a maximum likelihood approach (MLE) to obtain point estimates of the parameters (see Section 3.1.1). The Bayesian framework includes prior probability distributions for the BGMM parameters, such that the parameters behave stochastically. In this approach, the goal is to infer the posterior distribution of the parameters. Therefore, the Bayesian approach involves prior, likelihood and posterior functions, since the joint prior distribution is updated with data evidence to obtain the posterior distribution, via the Bayes' rule:

$$posterior = \frac{prior \times likelihood}{evidence} \tag{14}$$

In this framework, estimating the predicted distribution of target speech features ($\mathbf{y}^{(t)}$) requires marginalizing out the model parameters, since the parameters are modeled as random variables. Figure 4 shows a graphic model representation of a BGMM.

In the context of SSC, to estimate the posterior distribution parameters, we construct training data vectors $\boldsymbol{X}$ by concatenating a feature vector of source speech ($\boldsymbol{x}^{s)}$) and a feature vector of target speech ($\boldsymbol{x}^{(t)}$): $\boldsymbol{X} = [(\boldsymbol{x}^{(s)})^\intercal, (\boldsymbol{x}^{(t)})^\intercal]^\intercal$. Each sample ($\boldsymbol{X}_n$) has associated a latent observation ($\boldsymbol{z}_n$), which is a 1-of-$J$ binary vector with elements $z_{nj}$ for $j = 1, \ldots, J$. The number of Gaussian components is $J$, and $z_{nj} = 1$ if the observation belongs to the GMM's $j$th component, and 0 otherwise. That is, there is an underlying cluster-like structure in the data. We can represent the conditional distribution of latent variables $\boldsymbol{z}$ given weight coefficients $\boldsymbol{w}$ as:

$$p(\boldsymbol{z}|\boldsymbol{w}) = \prod_{j=1}^{J} w_j^{z_j}, \tag{15}$$

and the conditional distribution of the observed data, given the latent variables and model parameters, is of the form:

$$p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{X}|\boldsymbol{\mu}_j, \mathbf{\Sigma}_j)^{z_j}. \tag{16}$$

Thus, the likelihood is of the same form as in GMM (Equation 10), and we can

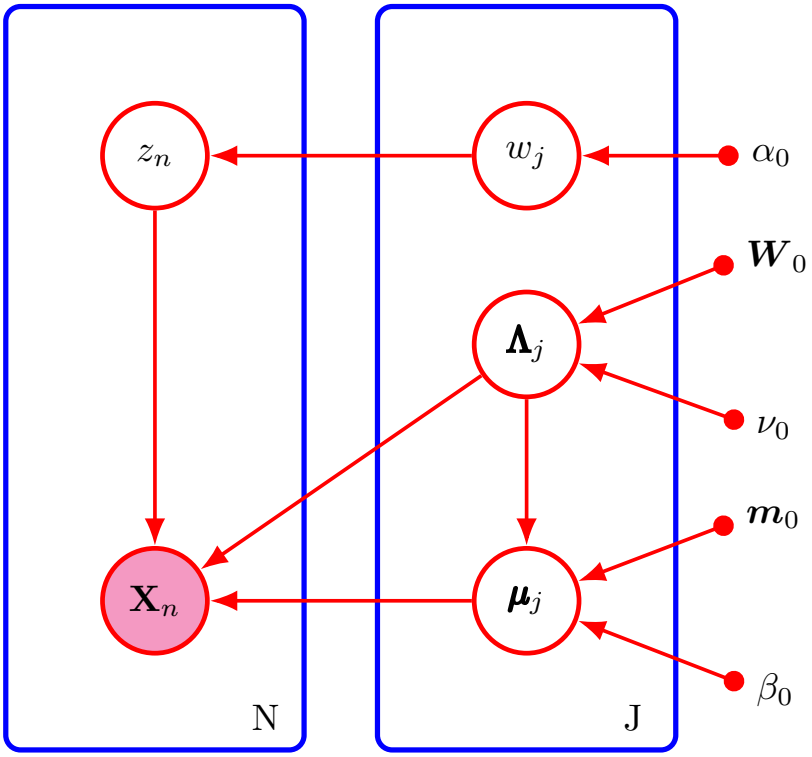


Figure 4: Graphical model of a BGMM, adapted from [86]. The nodes represent variables (and the node marked in pink is an observed variable). The red notches indicate point estimates of the hyper-parameters, and the arrows represent conditional dependencies. In addition, the two plates indicate repetition over the corresponding index, such that there is a set of $N$ i.i.d. observed data points $\{\boldsymbol{X}_n\}$, with corresponding latent points $\{z_n\}$, with $n = 1, 2, \ldots, N$, and parameters $\{w_j\}$, $\{\boldsymbol{\mu}_j\}$, and $\{\boldsymbol{\Lambda}_j\}$ of mixture components $j = 1, 2, \ldots, J$.

express it as:

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{j=1}^{J} w_j \mathcal{N}\left( \begin{bmatrix} \mathbf{x}^{(s)} \\ \mathbf{x}^{(t)} \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu}_j^{(s)} \\ \boldsymbol{\mu}_j^{(t)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_j^{(s,s)} & \boldsymbol{\Lambda}_j^{(s,t)} \\ \boldsymbol{\Lambda}_j^{(t,s)} & \boldsymbol{\Lambda}_j^{(t,t)} \end{bmatrix} \right), \quad \sum_{j=1}^{J} w_j = 1, \tag{17}$$

where now $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{w}\}$ represents the BGMM parameters[3], and $\boldsymbol{\mu} = \{\boldsymbol{\mu}_j\}$, $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_j\}$, and $\boldsymbol{w} = \{w_j\}$, for $j = 1, 2, \ldots, J$. $J$ is the number of mixture components, $w_j = p(j|\boldsymbol{\theta})$ is the prior probability of the $j$th component; and $\boldsymbol{\mu}_j$ and $\boldsymbol{\Lambda}_j$ are the

[3]In this thesis, we represent BGMM parameters with a different variable than the one used for GMM parameters, to emphasize the difference in nature between these two cases: GMM parameters ($\boldsymbol{\lambda}$) are fixed while BGMM parameters ($\boldsymbol{\theta}$) are random variables.

mean vector and precision matrix[4] of each $j$th component, respectively [74, 75].

Next, we need to select the prior probabilities of the model parameters, $\boldsymbol{\theta}$. For that, we have to take into account that the analysis is largely simplified if we choose conjugate priors. Thus, we select for mean and precision of each Gaussian component, a Normal-Wishart distribution:

$$p(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j) = p(\boldsymbol{\mu}_j|\boldsymbol{\Lambda}_j)p(\boldsymbol{\Lambda}_j) = \mathcal{N}(\boldsymbol{\mu}_j|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_j)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_j|\boldsymbol{W}_0, \nu_0), \tag{18}$$

where $\boldsymbol{m}_0$ defines the center, constant $\beta_0$ indicates how far the mean is on average from $\boldsymbol{m}_0$, $\boldsymbol{W}_0$ specifies the general shape of the distribution and $\nu_0$ is a constant that sets the variability of the data samples (that is, degrees of freedom)[88, 89]; $\nu_0 \geq F - 1$, where $F$ is the dimension of feature vectors in $\boldsymbol{X}$. Then, we select a Dirichlet prior distribution for the mixing coefficients:

$$p(\boldsymbol{w}) = Dir(\boldsymbol{w}|\boldsymbol{\alpha}_0), \tag{19}$$

where $\boldsymbol{\alpha}_0$ is a $J$-dimensional parameter. The hyper-parameters from these distributions (Eqs. 18-19) encode priori information about the data.

The joint distribution of all the variables is then:

$$p(\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{\theta}) = p(\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \tag{20}$$

We present this decomposition in graphical form in Figure 4.

Earlier in this section, we denoted $\boldsymbol{X}$ as the set of observed variables used in training; specifically in SSC, $\boldsymbol{X}$ is a set of concatenated feature vectors from source and target speech: $\boldsymbol{X} = [(\boldsymbol{x}^{(s)})^\intercal, (\boldsymbol{x}^{(t)})^\intercal]^\intercal$. Now we use $\boldsymbol{Z}$ to represent the set of all latent variables ($\boldsymbol{z}$) and model parameters ($\boldsymbol{\theta}$). The goal is to find first the model evidence, $p(\boldsymbol{X})$, and then to find the posterior distribution, $p(\boldsymbol{Z}|\boldsymbol{X}) = p(\boldsymbol{X}, \boldsymbol{Z})/p(\boldsymbol{X})$. However, there is no analytic solution for $p(\boldsymbol{Z}|\boldsymbol{X})$, since the extact inference of the true posterior $p(\boldsymbol{Z}|\boldsymbol{X})$ involves an intractable integration. Thus, in the work included in this thesis, we dealt with this intractability by performing an approximation of the posterior using *variational inference*. This method performs an exact inference of an approximate of the distribution of interest (in this case, the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X})$). The approximate distribution, $q(\boldsymbol{Z})$, will be a tractable distribution, and we denote it as *variational posterior*[5]. We achieve tractability by restricting the family of distributions $q(\boldsymbol{Z})$, while having at the same time as rich a family of approximating distributions $q(\boldsymbol{Z})$ as possible [89].

One of the approaches employed to restrict the family of approximating distributions $q(\boldsymbol{Z})$, and that we employ here, is by assuming that $q(\boldsymbol{Z})$ factorizes into $I$ disjoint groups (factors), while not making any further assumptions about the distributions. Then, we can represent the general form of $q(\boldsymbol{Z})$ as:

$$q(\boldsymbol{Z}) = \prod_{i=1}^{I} q_i(\boldsymbol{Z}_i). \tag{21}$$

[4]Henceforth we use precision $\boldsymbol{\Lambda}$ rather than covariance ($\boldsymbol{\Lambda} = (\boldsymbol{\Sigma})^{-1}$) in the representations, since it will simplify the mathematics of this section.

[5]Strictly, we should denote the variational posterior distribution as $q(\boldsymbol{Z}|\boldsymbol{X})$, but here we follow common convention of denoting it as $q(\boldsymbol{Z})$, for simplification.

The factorized form of variational inference belongs to the approximation framework denoted as *mean field theory* [90]. When applying this variational framework to the current case (a mixture of Gaussians), we assume the latent variables, $\boldsymbol{z}$, and the model parameters, $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$, to be conditionally (on $\boldsymbol{X}$) independent of each other. Thus, for the current case, the variational distribution $q(\boldsymbol{Z})$ factorizes between the latent variables and parameters as:

$$q(\boldsymbol{Z}) = q(\boldsymbol{z}, \boldsymbol{\theta}) \approx q(\boldsymbol{z})q(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Lambda}). \tag{22}$$

Furthermore, since we chose conjugate distributions for the priors, we can apply further factorization to $q(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$:

$$q(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{w}) \prod_{j=1}^{J} q(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j), \tag{23}$$

and we can observe a correspondence in functional form between the factors $q(\boldsymbol{z})$ and $q(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$, and their priors. Thus, $q(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j)$ follows Normal-Wishart distribution (as in Equation 18):

$$\hat{q}(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j) = \mathcal{N}(\boldsymbol{\mu}_j | \boldsymbol{m}_j, (\beta_j \boldsymbol{\Lambda}_j)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_j | \boldsymbol{W}_j, \nu_j), \tag{24}$$

and $q(\boldsymbol{w})$ follows Dirichlet distribution (as in Equation 19):

$$\hat{q}(\boldsymbol{w}) = Dir(\boldsymbol{w} | \boldsymbol{\alpha}). \tag{25}$$

In addition, factor $q(\boldsymbol{z})$ has the same functional form as that of the prior $p(\boldsymbol{z}|\boldsymbol{w})$ (Equation 15). Given all this, we achieve a practical, tractable solution for the posterior $p(\boldsymbol{Z}|\boldsymbol{X})$.

We obtain the functional form of the factors, $q(\boldsymbol{z})$ and $q(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ by optimizing $q(\boldsymbol{Z})$ using the Kullback-Leibler (KL) criteria: KL divergence between the true posterior $p(\boldsymbol{Z}|\boldsymbol{X})$ and variational posterior $q(\boldsymbol{Z})$ is minimized to find and estimate the factors of $q(\boldsymbol{Z})$. We use an iterative algorithm, as in the EM algorithm employed for GMMs, which alternates between two states that resemble the expectation (E) and maximization (M) steps of EM: 1) In the E-step of the variational case, the current estimate of distribution $q(\boldsymbol{\theta})$ is used to compute the responsibilities $r_j = \mathbb{E}(\boldsymbol{z}_j)$. 2) In the M-step, the responsibilities are fixed, and used to compute the variational posterior distribution over the parameters $\boldsymbol{\theta}$.

Once we have estimated the variational posterior using training data $\boldsymbol{X}$, we need to compute the posterior predictive density, $p(\boldsymbol{Y}|\boldsymbol{X})$, to use it at the conversion step for predicting the target feature $\boldsymbol{y}^{(t)}$. We obtain posterior predictive distribution $p(\boldsymbol{Y}|\boldsymbol{X})$ of sample $\boldsymbol{Y} = [\boldsymbol{y}^{(s)}, \boldsymbol{y}^{(t)}]^T$ given $\boldsymbol{X}$ by marginalizing the model parameters. The posterior predictive distribution has the form of a mixture of Student's t-distributions $St$ (more details in [89]):

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{\alpha'} \sum_{j=1}^{J} \alpha_j St(\boldsymbol{Y} | \boldsymbol{m}_j, \boldsymbol{L}_j, \boldsymbol{v}_j + 1 - F), \tag{26}$$

where $\boldsymbol{m}_j$ is the mean vector (given by Eq. 24) and $\boldsymbol{L}_j$ is the precision of the $j$th component; $F$ is the dimension of feature vectors in data set $\boldsymbol{X}$ (and $\boldsymbol{Y}$), $\boldsymbol{v}_j$ is given by Eq. 24, and the degrees of freedom for $j$th component is equal to $\boldsymbol{v}_j + 1 - F$. In addition, $\alpha_j$ is the $j$th element of vector $\boldsymbol{\alpha}$ (given by Eq. 25) and represents the mixture weight of the $j$th component, and $\alpha' = \sum_j \alpha_j$ [89]. Finally, we compute precision $\boldsymbol{L}_j$ as:

$$\boldsymbol{L}_j = \frac{(\boldsymbol{v}_j + 1 - F)\beta_j}{1 + \beta_j} \boldsymbol{W}_j, \tag{27}$$

where $\beta_j$ is given by Eq. 24; Eqs. 24 and 25 refer to the factors $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ and $q(\boldsymbol{w})$, respectively, obtained with variational inference.

Once we know the form of the posterior predictive distribution, and we express its parameters in matrix form as $\boldsymbol{m}_j = \begin{bmatrix} \boldsymbol{m}_j^{(s)} \\ \boldsymbol{m}_j^{(t)} \end{bmatrix}$ and $\boldsymbol{L_j} = \begin{bmatrix} \boldsymbol{L}_j^{(s,s)} & \boldsymbol{L}_j^{(s,t)} \\ \boldsymbol{L}_j^{(t,s)} & \boldsymbol{L}_j^{(t,t)} \end{bmatrix}$, then, the minimum mean square estimate (MMSE) of target feature vector $\boldsymbol{y}^{(t)}$ is given by:

$$\hat{\boldsymbol{y}}^{(t)} = \sum_{j=1}^{J} p(j|\boldsymbol{y}^{(s)}, \boldsymbol{X}, \boldsymbol{\theta}) \left[\boldsymbol{m}_j^{(t)} + \boldsymbol{C}_j(\boldsymbol{y}^{(s)} - \boldsymbol{m}_j^{(s)})\right], \tag{28}$$

where $p(j|\boldsymbol{y}^{(s)}, \boldsymbol{X}, \boldsymbol{\theta})$ is the marginal probability of the $j$th mixture component in Eq. 26, and we can express it as:

$$p(j|\boldsymbol{y}^{(s)}, \boldsymbol{X}, \boldsymbol{\theta}) = \frac{\alpha_j St(\boldsymbol{y}^{(s)}|\boldsymbol{m}_j^{(s)}, \boldsymbol{L}_j^{(s,s)}, \boldsymbol{v}_j + 1 - F)}{\sum_{j'}^{J} \alpha_{j'} St(\boldsymbol{y}^{(s)}|\boldsymbol{m}_{j'}^{(s)}, \boldsymbol{L}_{j'}^{(s,s)}, \boldsymbol{v}_{j'} + 1 - F)}, \tag{29}$$

and $\boldsymbol{C}_j$ is a linear transformation of $j$th mixture component [89, 88]:

$$\boldsymbol{C}_j = (\boldsymbol{L}_j^{(t,s)})^{-1} \boldsymbol{L}_j^{(s,s)}. \tag{30}$$

## 3.2 Data-driven, non-parallel mapping

Data collection of speech in different speaking styles tends to be very costly. Therefore, data sets consisting of speech produced using different speaking style are scarce, specially related to the parallel scenario described in Section 3.1. Related to this data scarcity problem, progress has recently been made to develop SSC techniques based on non-parallel scenarios.

In [91], cycle-consistent generative adversarial networks (cycleGANs) [92] were used to convert normal speech to Lombard speech (and vice-verse). CycleGANs learn a bidirectional deterministic mapping between two domains, in this case normal and Lombard speech, using non-parallel training data from both domains. The cycleGAN-based mapping approach has been used in VC, in which non-parallel conversion approaches have been studied recently. CycleGAN is a recent alternative to the most common non-parallel mapping approach used in VC, the technique called Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment method (INCA) [93]. The main advantage of cycleGANs over INCA is that

cycleGANs do not need frame alignment (which is not a trivial task) before model training. INCA was also employed in [91] for comparison, and in overall cycleGANs proved to give better SSC performance in terms of strength in the perceptual change between the two speaking styles, and also in terms of speech quality. The data employed in [91] was a corpus of read and conversational Lombard speech, along with read speech uttered in normal speaking style, all in Finnish language [94].

Another SSC study based on non-parallel mapping was presented in [95]. This study proposed an extension of cycleGANs: augmented cycleGANs [96], which improved over the limitations present in cycleGANs. The main problem with cycleGANs is that they learn deterministic mappings from the training data. In the augmented cycleGAN approach, mappings are defined over augmented latent spaces such that the augmented cycleGAN model learns many-to-many bidirectional mappings between two domains (in SSC, the source and target speaking style). The data set employed in [95] consisted of: 1) the same corpus employed in [91], which was uttered in Finnish [94], and 2) a corpus of English, read utterances that contain both normal and Lombard speaking styles [97].

# 4 SSC from normal speech to high vocal effort speech

SSC refers to the technology to convert speech from its source speaking style to a given target style by keeping the linguistic content and the speaker identity unchanged. It is desirable that the perceptual quality of speech could be maintained in the SSC process and the converted output sample would sound as natural as the input sample.

The review we present in this section focuses on SSC from normal speaking style to speaking styles of high vocal effort. While VC and SPSS have been studied extensively, there are clearly less studies in SSC. In vocal-effort-focused SSC, studies investigating conversion of whispered speech to normal speech have been most common (e.g. [98, 99, 100, 101, 102, 103, 104, 105, 106, 107]). In comparison, there is little research in conversion to speaking styles of high vocal effort. To the best of our knowledge, the topic has been previously studied only in [11, 9, 72, 10, 27, 73, 76, 91, 95]. Some of these studies investigate only limited data such as conversion of single words [72] or logatomes (which are pseudo-words of one or many syllables) [10]. In addition, in some studies, such as [11], only a few sentences were converted. In contrast, the studies reported in [9, 27, 73, 76, 91, 95] were performed with more realistic conditions, since the conversion involved a data set of full speech sentences, rather than converting smaller speech units. The speech samples used in [91, 95] were particularly realistic, since these studies included not only read speech but also conversational speech.

In this section, we present SSC methods divided into two categories, based on the underlying signal processing methodology. 1) The vocoder-based parametric approach uses a vocoder for feature extraction and synthesis. 2) The direct transformation approach makes the conversion by directly filtering the speech signal (either in the time or frequency domain). In addition, this section includes a separate discussion of applying SSC in speaker recognition under vocal effort mismatch.

## 4.1 Vocoder-based parametric SSC approaches

The widely used STRAIGHT vocoder was adopted in a SSC system for feature extraction and synthesis in [72]. This study investigated style conversion from single words of normal speech to the corresponding units of Lombard speech. A statistical study of the differences between normal and Lombard speech was performed using the SUSAS database [108]. Based on the values extracted, scaling of $f_0$, spectral envelope and phone duration was conducted. In processing the spectrum, formant frequencies and the distribution of spectral energy were modified. The converted isolated words were evaluated using listening tests on naturalness, similarity and voice quality. The tests were run in a manner that allowed evaluating both the individual modifications and the combined modifications. The results showed that by modifying solely the individual features did not yield Lombard-like speech. In contrast, by modifying all the selected features, Lombard-like speech was obtained.

The study reported in [9] investigated normal-to-Lombard conversion of synthetic speech to improve its intelligibility. Two different synthetic voices were included in

the study: a unit selection voice and a diphone voice. For conversion, cepstral vectors were converted using a ML-based mapping with a GMM. The converted speech sample was synthesized using a vocoder. The intelligibility improvement was evaluated using the word error rate (WER) measured in a subjective listening test by presenting the stimuli both with and without background noise. The converted diphone voice showed an improvement in WER over the non-converted counterpart. In contrast, the converted unit selection voice did not introduce any improvement in WER over the unmodified version. The authors of [9] argued that this may have been due to the degradation of speech quality of converted speech in the latter case, which lead to the loss of the advantage obtained by the conversion.

In [27], a normal-to-Lombard SSC task was studied by using a BGMM-based ML technique to map some of the features ($f_0$, spectral tilt, and energy) computed from normal speech to the corresponding features of Lombard speech. In addition, the utterance duration was also modified by using the cubic spline interpolation at the frame-level for every feature. The modified features were then employed by the vocoder to synthesize Lombard-like speech. In this study, two vocoders were used for comparison: STRAIGHT, and a glottal vocoder that is a variant of the vocoder presented in [28]. For evaluation, subjective listening tests on naturalness, and similarity were conducted. The results showed that for both vocoders, conversion of normal speech to Lombard speech was achieved. However, the naturalness of converted speech was clearly higher when the glottal vocoder was used. This study corresponds to the second topic of this thesis, which is the conversion of normal speech to Lombard speech. Thus, more details can be found in Sections 7 and 8.

The study reported in [76] is an extension of the work published in [27]. In [76], the studied task was also normal-to-Lombard conversion. Three vocoders were compared: STRAIGHT, GlottDNN, and PML. Regarding the features, the same ones as in [27] were modified ($f_0$, spectral tilt, energy and duration), though in this case the features were extended using adjacent frames. All features except duration were mapped using three different ML-based techniques for comparison: a GMM, a BGMM, and a feed-forward DNN. As in [27], duration was transformed using the cubic spline interpolation. The conversion results were evaluated in listening tests that evaluated the similarity of the converted samples to Lombard speech, and speech quality. In addition, intelligibility was evaluated using an objective measure called speech intelligibility in bits (SIIB). The results of this study showed that while Lombard conversion was achieved by the proposed system, there was a trade-off between the Lombardness achieved and the quality of converted speech. In terms of quality, GlottDNN proved to be the best, while speech converted using PML showed a larger amount of Lombardness perceived. Of all the possible combinations in the vocoder and mapping techniques, PML with GMM seemed to give a better compromise between quality and Lombardness. The SIIB measure showed the largest improvement in speech intelligibility in background noise when DNN mapping was used with STRAIGHT or PML.

The study reported in [91] employed a non-parallel ML-based technique, cycle-GAN [92], in a normal-to-Lombard SSC task. This differs from the aforementioned studies employing ML-based mapping [9, 27, 76], which require parallel training

data. In addition to cycleGAN, the study also included the GMM and INCA algorithms for comparison. On the other hand, the PML vocoder was used in the system for feature extraction and synthesis. The features mapped were $f_0$, voicing decisions (V/UV), and the spectral envelope. The duration was also modified using the cubic spline interpolation. The mapping techniques were compared in listening tests by evaluating the similarity between the converted samples and natural Lombard speech, and speech quality. CycleGAN showed better performance than the other techniques in the two attributes evaluated.

In [95], an extension to the study conducted by the same authors in [91] was proposed by investigating again conversion of normal speech to Lombard speech. The augmented cycleGAN technique proposed for mapping is an extension of the cycleGAN mapping technique used in [91]. The cycleGAN was used in this study as a reference mapping technique. The vocoder used was the same as in [91] (PML), and the mapped features were $f_0$ and spectral envelope. The evaluations of the speech quality and intelligibility showed that the proposed system was able to achieve Lombard-like speech conversion. However, the improvement of intelligibility was only shown by the SIIB measure, but not in the subjective listening tests. The authors argued that this might be due to the degradation caused by the vocoding and mapping operations.

## 4.2 Direct transformation SSC approaches

In [11], a SSC system was proposed for conversion to different vocal efforts. In this work, glottal source was firstly extracted using the iterative adaptive inverse filtering (IAIF) method [109] on a frame-basis. Then, the algorithm in [110] was used to decompose the glottal source signal into periodic and aperiodic components, and the spectral content of these two components was modified with time-varying linear filtering. This type of filtering is achieved by: 1) multiplying the discrete Fourier transform (DFT) samples (in case of [11], the periodic and aperiodic components) by the frequency samples of the filter used, and 2) performing overlap-add (OLA) on the modified signal (in case of [11], a weighted sum of the modified periodic and aperiodic components) [111]. Finally, the glottal source signal synthesized with OLA is filtered by a vocal tract filter, to obtain the converted speech signal. In this study, only a few speech sounds were converted as examples by studying target styles whose vocal effort was both lower and higher compared to the vocal effort of the source style. In addition, no subjective or objective evaluations of the converted sounds were conducted.

The study reported in [10] focused on conversion of logatomes of normal speech to sounds of lower and higher vocal effort. A system was proposed, inspired by the adaptive pre-emphasis linear prediction (APLP) method published in [112] that used APLP for transforming high vocal effort speech to breathy speech. In [10], the harmonics plus noise model (HNM) was used to model speech as a sum of: 1) a harmonic (pseudo-periodic) component, which is characterised by amplitude, frequency, and phase vectors, and 2) a noise (aperiodic) component. The HNM model had previously been shown to give good quality in prodosy modification

[113]. Harmonic and noise components were extracted by LP analysis and by using the algorithm proposed in [114]. Thus, as in the case of [11], the conversion in [10] is based on a periodic/aperiodic decomposition, though in [10] the signal to be decomposed is speech, while in [11] it is the glottal source signal. Using parallel data of the HNM parameters from the source and target speech signals, source speech is time-aligned to target speech. Then, $f_0$, spectral emphasis, and energy are modified to match the same features of the target speech; spectral emphasis was modified using only the harmonic component of the signal. For all of these transformations, the modification of the spectral emphasis produces larger changes in terms of vocal effort. The spectral emphasis transformation was originally proposed in APLP, but using LP residuals rather than HNM parametrization. The proposed system was evaluated in subjective listening tests, and the evaluation included the APLP method as a reference. The results showed that while the proposed approach is able to convert speech of the source style to the target speaking style, there is a trade-off between the level of conversion and the quality achieved in the conversion. That is, the larger the perceptual effect of the conversion, the larger was the degradation of the speech quality. This study also proved that APLP can be used also for conversion to styles of higher vocal effort.

Finally, the study presented in [73] investigated SSC of normal speech to Lombard speech with the application to improve speech intelligibility in noisy car environment. To achieve this, modifications of three aspects of speech were studied: 1) time modification using the transformation of $f_0$ marks based on scaling, and applying pitch-synchronous overlap-add (PSOLA) with these new $f_0$ marks, 2) smoothed shifting of formant frequencies for voiced segments such that formants are shifted away from the regions where noise is strong, 3) energy redistribution between voiced and unvoiced segments (specifically, unvoiced regions were boosted to become more intelligible under noisy conditions). For evaluation of intelligibility, a subjective listening test was run using the converted Lombard-like speech utterances corrupted with additive car noise at different signal-to-noise ratios (SNRs). The converted samples presented in some cases the combination of all the three proposed modifications, and in other cases only individual modifications were applied. The subjective evaluations included both normal listeners and listeners with hearing impairments. Intelligibility was also evaluated using several objective measures (e.g., speech intelligibility index (SII) and mutual information (MI)). The results of both the subjective and objective tests showed that the proposed speech modifications yielded more intelligible speech. The best results were obtained when all the proposed modifications were applied.

## 4.3 SSC for speaker recognition under vocal effort mismatch

Sections 4.1 and 4.2 omitted mentioning vocal-effort-related SSC studies in which the speech conversion task has a specific application in computer-based speech technology, such as in a ASR system or in a automatic speaker recognition system. In these cases, subjective quality of the converted speech samples is of no interest. Thus, the converted samples are not required to retain high speech quality, but

rather it is desirable that the attributes of the converted samples are well suited for the application at hand (for example, good speaker-discriminating power, in case of using the converted samples in automatic speaker recognition). We could regard such techniques as a subgroup of SSC methods, given that their goal is very different from the rest of applications of SSC. Here the focus is specifically on the application of SSC in speaker recognition under vocal effort mismatch, since one of the topics in this thesis (first topic, published previously in [22]) represents this subgroup.

Many of the variabilities found in speech recordings belonging to the same speaker are a problem for speaker recognizer systems, since they cause a mismatch between training and test samples. These speaker variabilities can origin both from extrinsic factors to the speaker (such as the acoustic environment or the transmission channel) and intrinsic factors (such as the vocal effort or the age). While vocal effort mismatch has proved to affect considerably the speaker identification rates [19], this mismatch problem has not been covered in research studies as much as, for example, the impact of background noise.

To tackle the vocal effort mismatch problem in speaker recognition, one of the ways employed is processing the system's features [115, 21, 116, 117, 118, 119, 120, 23]. We can divide these feature-based techniques into three different approaches: 1) methods in which the louder speech and the softer speech[6] are processed such that a middle point between the two is found, for which the difference in acoustic features between these two signals is reduced [115, 21, 116, 117, 120]; 2) methods in which the louder speech is processed such that its features are closer to those of the softer speech [118, 119]; 3) the vice versa situation of the second case. That is, approaches in which the softer speech is processed and its features approach those of the louder speech [115, 119, 22, 23].

From all the above studies on vocal effort mismatch in speaker recognition, we can only consider [22] to belong to SSC: it is the only study in which the processing involves an acoustic-to-acoustic conversion from the source style to the target style. While the method proposed in [22] was originally intended for SSC in general, for the current application the conversion remains only at signal frame-level. This is because the full target speech signal does not need to be synthesized given that the converted speech frames are fed to the speaker recognition system. The SSC system proposed in [22] is based on a direct transformation approach. The proposed system employs the PSM-GMM method to transform the Mel-scale filter bank energies (MFBEs) of normal speech (from enrollment utterances) towards their counterparts in shouted speech. We mentioned PSM-GMM earlier in Section 1: it combines a signal processing technique with a ML-based mapping technique to do the transformations on the speech signal. The study carried out in [22] corresponds to the first topic presented in this thesis. Thus, we present more details of this study in Sections 5 and 6.

[6]For concision, here *louder speech* refers to the speech signal with a higher vocal effort out of the two speech signals available. Likewise, *softer speech* refers to the speech signal with a lower vocal effort out of the two available.

# 5 PSM–GMM: A direct transformation SSC system

This section introduces the theoretical foundation of the SSC method proposed in the section on conversion of normal speech to shouted speech. This method works directly on the source speech signal to perform the task of converting it to the target source speech signal (that is, it is a direct transformation SSC approach).

The proposed system processes the speech signal using a novel technique called PSM. This technique adjusts the MFBEs of the source speech (in this case, normal speech) to approach those of the target speech (here, shouted speech). The spectral adjustment uses PSM to estimate a filter, which we henceforth call the *mapping filter*, and applies it to unprocessed normal speech. This adjusts the normal speech's MFBEs to approach those of shouted speech. To estimate the mapping filter, PSM requires pairs of utterances from both the source (normal speech) and target (shouted speech). To avoid the need for parallel data in the conversion, we use statistical mapping in conjunction with PSM. Specifically, we train a statistical model to learn the dependencies between features of normal and shouted speech. Once trained, the model can estimate the mapping filter required to convert a normal speech sample frame to shouted speech. In this study, the statistical model employed was GMM. This model has been used frequently in voice conversion tasks [74]. More details on GMMs can be found in Section 3.1.1.

The overall system proposed for the conversion of source speech to target speech is denoted PSM–GMM, as these two parts constitute the system. We present the signal processing technique PSM in Section 5.1. Finally, we introduce the proposed system in Section 5.2.

## 5.1 PSM

PSM was first proposed in [121] as a novel approach to speech synthesis for perceptual scale-based spectral matching of a synthetic speech signal to the target natural speech signal. Motivated by the PSM design presented in [121], our study aims to find an all-pole mapping filter $H_m(z)$ with impulse response $h_m(k)$ such that the mel spectrum $s_{source}(k) * h_m(k)$ matches that of $s_{target}(k)$, where $s_{source}(k)$ is the frame of the speech signal in the source speaking style, and $s_{target}(k)$ is the corresponding speech frame uttered in the target speaking style. We estimate the mapping filter by minimizing the distance between the MFBEs of the source and target speech signals.

Before going into details about the stages of the PSM technique, we note that in the past, both power and magnitude have been employed to obtain the MFBEs. In our study, we employ the power spectrum. MFBEs are usually computed by applying a mel-scale filter bank to the fast Fourier transform (FFT) power spectrum of the speech frame at hand. However, in this work, we use the same computation procedure as in [121], which functions equivalently despite being different from the standard approach. This alternative procedure involves two steps: First, we compute a mel-warped spectrum of the given speech frame by performing mel scale-based spectral interpolation of the FFT power spectrum bins, as in [122]. Then,

we pass the mel-warped spectrum through a uniform-scale triangular filter bank $\boldsymbol{T} = [\boldsymbol{t}_0^T, \boldsymbol{t}_1^T, \cdots, \boldsymbol{t}_{M-1}^T]^T$ with $M$ filters $\boldsymbol{t}_i$ $(0 \leq i \leq M-1)$ of equal passband width and 50% overlap.

A flowchart showing the stages of the PSM technique appears in Figure 5. In PSM, we first compute the MFBE vector (with dimension $M \times 1$) of the target speech $s_{target}(k)$ as

$$\boldsymbol{\mathcal{E}}_{target} = \boldsymbol{T}|S_{target}(\tilde{\Omega})|^2, \tag{31}$$

where $|S_{target}(\tilde{\Omega})|^2$ is the mel-warped power spectrum of $s_{target}(k)$, and $\tilde{\Omega}$ is the index of the warped FFT bins [122]. The following power spectrum references are defined in the frequency-warped domain.

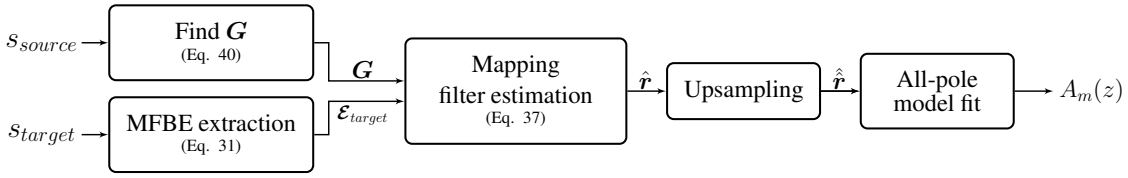


Figure 5: Block diagram of the stages of the perceptual spectral matching (PSM) technique. MFBE stands for mel-scale filter bank energy. $H_m(z) = 1/A_m(z)$ is the final $p$th-order, all-pole mapping filter.

In general, there is no unique inverse transformation to return from the MFBE vector associated with the mapping filter's power spectrum to the corresponding full-length spectrum (with $N_{FFT}$ points) due to the dimensionality reduction caused by computing the filter bank energies. However, a unique solution can be obtained if it is assumed that the (mel-warped) power spectrum of the mapping filter, denoted $|H_m(\tilde{\Omega})|^2$ , is piecewise constant (see Figure 6). In this case, the $i$th segment of the power spectrum takes the value $r_i$, with the $M \times 1$ vector $\boldsymbol{r}$ containing the values of all the segments. $\boldsymbol{r}$ is dubbed the *elementary power spectrum*.

The boundaries between segments of the power spectrum are determined by the uniform-scale, triangular filter bank $\boldsymbol{T}$. The $i$th segment (with value $r_i$) is limited to the region where filter $\boldsymbol{t_i}$ in $\boldsymbol{T}$ has a larger spectral amplitude than its neighbouring filters $\boldsymbol{t_{i-1}}$ and $\boldsymbol{t_{i+1}}$. Such construction of $\boldsymbol{r}$ partitions each filter $\boldsymbol{t_i}$ into three different regions. In each region, the output of filter $\boldsymbol{t_i}$ is affected by a different segment of $\boldsymbol{r}$: In the central region $R_i^c$, filter $\boldsymbol{t_i}$ has the highest spectral amplitude, and the contributing segment to the filter output is $r_i$. In the lower region $R_i^l$, filter $\boldsymbol{t_{i-1}}$ has the largest amplitude, and the contributing segment is $r_{i-1}$. In the upper region $R_i^h$, filter $\boldsymbol{t_{i+1}}$ has the largest amplitude, and $r_{i+1}$ contributes to the filter output. Based on this construction, the matching filter power spectrum $|H_m(\tilde{\Omega})|^2$ can be expressed as

$$|H_m(\tilde{\Omega})|^2 = r_i \text{ when } \tilde{\Omega} \in R_i^c. \tag{32}$$

Given this configuration, we can identify the outputs of the filter bank, which correspond to the MFBE vector $\hat{\boldsymbol{\mathcal{E}}}_{target}$ of the speech that is mapped from the original

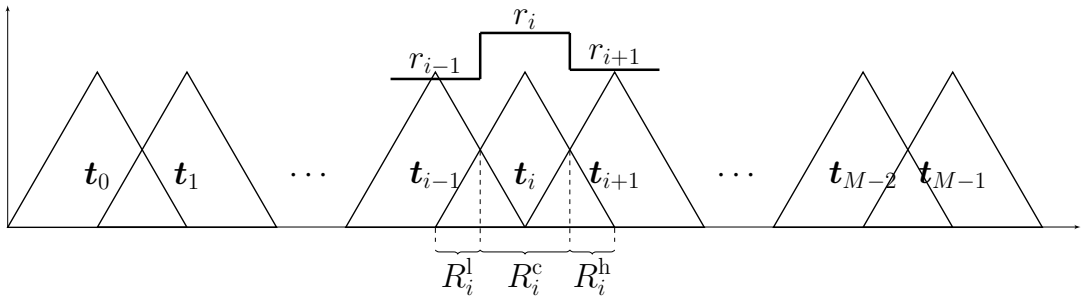


Figure 6: Piecewise constant upsampling of $\boldsymbol{r}$, which is the elementary power spectrum of the mapping filter. Note that the triangular filter banks are of equal width due to mel-warping of the input signal.

speech to shouted speech. We refer to the resulting signal as target-like speech. The $i$th element of the MFBE vector $\hat{\boldsymbol{\mathcal{E}}}_{target}$ of the target-like speech frame can be expressed as

$$\begin{aligned}\hat{\mathcal{E}}_{target,i} = &\sum_{\tilde{\Omega}\in R_i^l} t_i(\tilde{\Omega}) \cdot |S_{source}(\tilde{\Omega})|^2 \cdot r_{i-1} \\ &+ \sum_{\tilde{\Omega}\in R_i^c} t_i(\tilde{\Omega}) \cdot |S_{source}(\tilde{\Omega})|^2 \cdot r_i \\ &+ \sum_{\tilde{\Omega}\in R_i^h} t_i(\tilde{\Omega}) \cdot |S_{source}(\tilde{\Omega})|^2 \cdot r_{i+1},\end{aligned} \tag{33}$$

where $|S_{source}(\tilde{\Omega})|^2$ is the mel-warped FFT power spectrum of the source speech. Given that $r_i$ is independent of index $\tilde{\Omega}$, Eq. 33 can be rewritten as

$$\hat{\mathcal{E}}_{target,i} = G_i^l \cdot r_{i-1} + G_i^c \cdot r_i + G_i^h \cdot r_{i+1}, \tag{34}$$

where $G_i^l$, $G_i^c$ and $G_i^h$ are computed as

$$\begin{aligned}G_i^l &= \sum_{\tilde{\Omega}\in R_i^l} t_i(\tilde{\Omega}) \cdot |S_{source}(\tilde{\Omega})|^2 \\ G_i^c &= \sum_{\tilde{\Omega}\in R_i^c} t_i(\tilde{\Omega}) \cdot |S_{source}(\tilde{\Omega})|^2 \\ G_i^h &= \sum_{\tilde{\Omega}\in R_i^h} t_i(\tilde{\Omega}) \cdot |S_{source}(\tilde{\Omega})|^2.\end{aligned} \tag{35}$$

Writing the output of all the filters in matrix form leads to the following equation:

$$\underbrace{\begin{bmatrix} G_0^c & G_0^h & 0 & \dots & & 0 \\ G_1^l & G_1^c & G_1^h & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ & & & G_{M-2}^l & G_{M-2}^c & G_{M-2}^h \\ 0 & \cdots & & 0 & G_{M-1}^l & G_{M-1}^c \end{bmatrix}}_{\boldsymbol{G}} \underbrace{\begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{M-2} \\ r_{M-1} \end{bmatrix}}_{\boldsymbol{r}} = \underbrace{\begin{bmatrix} \hat{\boldsymbol{\varepsilon}}_{target,0} \\ \hat{\boldsymbol{\varepsilon}}_{target,1} \\ \vdots \\ \hat{\boldsymbol{\varepsilon}}_{target,M-2} \\ \hat{\boldsymbol{\varepsilon}}_{target,M-1} \end{bmatrix}}_{\hat{\boldsymbol{\varepsilon}}_{target}}. \quad (36)$$

In order to find $\boldsymbol{r}$, one can use the target speech energies $\boldsymbol{\varepsilon}_{target}$ as $\boldsymbol{G}\boldsymbol{r} = \boldsymbol{\varepsilon}_{target}$. However, it is not possible to solve directly from the expression $\boldsymbol{r} = \boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_{target}$, as doing so could yield negative values for the elementary power spectrum $\boldsymbol{r}$. Therefore, we instead formulate the mapping filter as a non-negative least square (NNLS) problem [123], where the objective function to be minimized is $(\hat{\boldsymbol{\varepsilon}}_{target} - \boldsymbol{\varepsilon}_{target})^2 = (\boldsymbol{G}\boldsymbol{r} - \boldsymbol{\varepsilon}_{target})^2$, given the element-wise non-negativity constraint $r_i \geq 0$. This can be expressed as

$$\hat{\boldsymbol{r}} = \arg\min_{r_i \geq 0}\{(\boldsymbol{G}\boldsymbol{r} - \boldsymbol{\varepsilon}_{target})^2\}. \quad (37)$$

We solve this equation with the classical NNLS algorithm, which ensures that convergence is reached when the pseudo-inverse of $\boldsymbol{G}$ is well defined [123]. Additionally, it should be noted that even though the classical NNLS algorithm is computationally expensive for large-scale problems due to the matrix pseudo-inverse computation, it is suitable for small problems like the one handled here.

Once we obtain an estimate for $\hat{\boldsymbol{r}}$, we upsample it in order to obtain a full-length $N_{FFT}$-point representation in the form of $\hat{\hat{\boldsymbol{r}}}$ before fitting an autoregressive model $H_m(z)$ of the mapping filter, where $|H_m(\tilde{\Omega})|^2 = \hat{\hat{\boldsymbol{r}}}$. We perform upsampling by using an expansion matrix $\boldsymbol{E}$ to expand the $M \times 1$ vector $\hat{\boldsymbol{r}}$ into a full-length power spectrum such that $\hat{\hat{\boldsymbol{r}}} = \boldsymbol{E}\hat{\boldsymbol{r}}$. This leads to a convenient matrix expression for the MFBE vector of the target-like speech frame:

$$\hat{\boldsymbol{\varepsilon}}_{target} = \underbrace{\boldsymbol{T}\boldsymbol{D}\boldsymbol{E}}_{=\boldsymbol{G}}\hat{\boldsymbol{r}}, \quad (38)$$

where $\boldsymbol{D}$ is a diagonal matrix with the vector elements of the mel-warped target speech power spectrum $|H_m(\tilde{\Omega})|^2$ on its diagonal.

Matrix $\boldsymbol{E}$ is constructed based on the assumed shape of the mapping filter's power spectrum, while the elements of vector $\boldsymbol{r}$ can be understood as points sampled from $H_m$ at the triangular filter centres. Thus, $\boldsymbol{E}$ can be chosen to perform interpolation of the elements of $\boldsymbol{r}$. The piecewise constant assumption corresponds to doing nearest-value (0th order) interpolation between $\boldsymbol{r}$'s samples. In such a case,

the expansion matrix $\boldsymbol{E}$ is

$$\boldsymbol{E}(i,k) = \begin{cases} 1, \text{ when } k \in R_i^c \\ 0, \text{ otherwise.} \end{cases} \tag{39}$$

If matrix $\boldsymbol{E}$ is defined as in Eq. 39, $\boldsymbol{G} = \boldsymbol{TDE}$ is equivalent to the form of $\boldsymbol{G}$ presented in Eq. 37. However, in this study, we employ first-order interpolation for expansion in order to obtain a piecewise linear spectrum. This is achieved by using expansion matrix $\boldsymbol{E} = \boldsymbol{T}^T$ and in turn solving Eq. 37 with

$$\boldsymbol{G} = \boldsymbol{TDT}^T. \tag{40}$$

Finally, once we have computed the full-length spectrum, we obtain from it the mel-warped, all-pole mapping filter $H_m(z) = 1/A_m(z)$ (with impulse response $h_m(k)$). For this, we calculate the autocorrelation from $\hat{\hat{\boldsymbol{r}}}$ and then utilize Levinson–Durbin recursion [124, 125].

## 5.2 PSM–GMM algorithm

The proposed algorithm uses GMM statistical mapping to carry out automatic spectral mapping of source speech to target speech. Thus, using the trained GMM, the statistical mapping makes it possible to predict the mapping filter that corresponds to a given frame of source speech in order to transform it into shouted speech. A block diagram of the PSM-GMM algorithm is presented in Figure 7. Note that while this algorithm was earlier presented in the context of the first topic (conversion of normal speech into shouted speech), the following figure presents it in a generalized form that does not possess a specific source style, a specific target style, or a specific feature vector from the source speech. This is because the algorithm may be applied to other style conversions.

In this conversion system, the statistical model must first be trained to perform automatic adjustment of the mel spectral band energies of source speech to target speech during the conversion stage. The processing occurs at frame level in this system, meaning that first a frame of source speech $s_{source}(k)$ and a frame of target speech $s_{target}(k)$ are extracted. The frames of these speech utterances must be aligned prior to this, since the duration of utterances sometimes differs across speaking styles. The $s_{source}(k)$ and $s_{target}(k)$ frames are then fed to the PSM algorithm, which computes a mapping filter $H_m(z) = 1/A_m(z)$. Next, pairs of feature vectors are extracted (corresponding to the source–target training frame pairs) and are concatenated to be used for training the GMM. We extract the feature vector $\boldsymbol{x}^{(s)}$ that corresponds to the frame $s_{source}(k)$. In the case of the target feature vector ($\boldsymbol{x}^{(t)}$), the mapping filter $A_m(z)$, extracted via PSM based on frames $s_{source}(k)$ and $s_{target}(k)$, is actually employed as vector.

Once the GMM has been trained, automatic spectral mapping can be performed. This is done by first extracting feature vector $\boldsymbol{y}^{(s)}$ from source speech

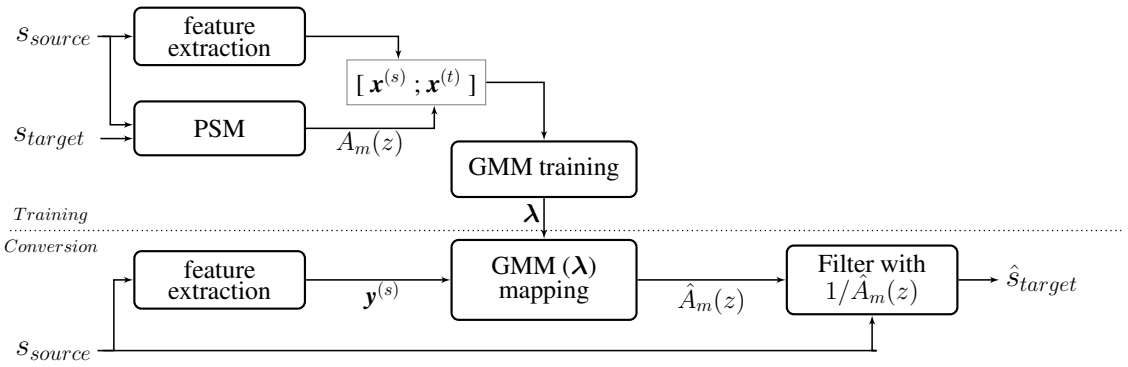


Figure 7: Flowchart of the PSM–GMM algorithm for converting frames of source speech to target-like speech by employing Eqs. 11–13. To train the GMM, we employed concatenated feature vectors, which were built 1) from source speech feature vectors $\boldsymbol{x}^{(s)}$ and 2) from feature vectors $\boldsymbol{x}^{(t)}$ computed from the mapping filters $H(z) = 1/A(z)$ obtained with PSM. The estimated target speech frame at the output $\hat{s}_{target}$ was produced by filtering a source speech frame $s_{source}$ with the predicted mapping filter $\hat{H}(z) = 1/\hat{A}(z)$.

frame $s_{source}(k)$[7] and feeding it to the trained GMM. Then, the GMM uses $\boldsymbol{y}^{(s)}$ to get an estimate of $\boldsymbol{y}^{(t)}$ (denoted $\hat{\boldsymbol{y}}^{(t)}$), which is the filter $\hat{A}_m(z)$. Finally, $s_{source}(k)$ is filtered with the estimated mapping filter $\hat{H}_m(z) = 1/\hat{A}_m(z)$, and an estimate of the target speech frame $\hat{s}_{target}(k)$ is obtained.

[7]Here and henceforth, we follow the same convention employed in Section 3, and denote the training feature vectors of source and target speech as $\boldsymbol{x}^{(s)}$ and $\boldsymbol{x}^{(t)}$ respectively, while $\boldsymbol{y}^{(s)}$ and $\boldsymbol{y}^{(t)}$ are the corresponding feature vectors in the conversion stage.

# 6 Experimental work (topic I): Normal-to-shouted speech, PSM–GMM-based SSC with application to speaker recognition under vocal effort mismatch

The experimental work in the first topic of this thesis (conversion of normal speech to shouted speech) dealt with the application of the PSM–GMM algorithm presented in Section 5 to a speaker recognition task in which there was a mismatch between test and enrollment utterances. In particular, the enrollment utterances used in the speaker recognition system were uttered as normal speech, while the test utterances were shouted. Our approach aims to minimize the spectral mismatch between such utterances by converting the MFBEs of normal speech (in enrollment) into their counterparts in shouted speech (in test). The goal of this procedure is to improve speaker recognition performance in the aforementioned mismatch conditions.

## 6.1 Data

In this experimental work, the same speech data set was used in the evaluation of the speaker recognition system and at the PSM–GMM processing step for training the GMM employed in mapping. This data set includes speech recordings from 22 Finnish speakers, half of which are female and half of which are male [71, 126]. The recordings originally had a sampling rate of 16 kHz, but we downsampled them to 8 kHz for out experiment. The data set contains 24 short utterances, about two seconds each, enunciated by each speaker. Each utterance was recorded in two different speaking styles: normal and shouting. The recordings were made in an anechoic chamber. Prior to recording, the speakers were not instructed to aim for any particular SPL. They were instead told to enunciate their sentences with great vocal effort for the shouted speech recordings. The difference in SPL between the two speaking styles varied greatly among speakers: it ranged from 17 to 28 dB for female speakers and from 15 to 33 dB for male speakers.

The speaker recognition system employed in this experimental work was developed in [127]. As we did not use its training data set in our experiments, we do not present that information here, but we explain it below in Section 6.2.1.

## 6.2 Experimental setup

### 6.2.1 Speaker recognition system

Our research employs a state-of-the-art, *i*-vector-based speaker recognition system developed at Radboud University Nijmegen for a submission to the 2012 NIST Speaker Recognition Evaluation [127]. In that work, feature vectors were extracted as input to the speaker recognition system. First, the speech of each recording was split in Hamming-windowed frames of 30 ms, with a frame shift of 15 ms. All frames (voiced and unvoiced) of the utterances were retained during both the enrollment

and recognition phases, since the duration of the utterances was very short. For each frame, a 60-dimensional feature vector was obtained, and it consisted of the following:

1. The first 20 Mel-frequency cepstral coefficients (MFCCs) [128], except for the first one ($c_0$). The computation of MFCCs requires a power spectrum estimate, and this was obtained via 12th-order LP analysis. Additionally, RASTA filtering [129] was applied to the MFCCs.

2. Frame energy.

3. Dynamic $\Delta$ and $\Delta\Delta$ features, computed from the vector resulting from the concatenation of features 1) and 2).

In the $i$-vector-based speaker recognition system [127], a gender-dependent universal background model (UBM) of 512 components was trained using a subset of the NIST SRE 2004, Fisher, Callfriend, and Switchboard speech data sets. For each utterance of interest, sufficient statistics were computed. The total variability space was trained using the NIST SRE 2004–2008, Fisher, and Switchboard corpora, from which utterance-level, 450-dimensional $i$-vectors were then computed. Next, the $i$-vectors were post-processed by using linear discriminant analysis (LDA) to project the vectors onto a 200-dimensional space such that their dimensionality was reduced and the separability between speakers was increased. The $i$-vectors were further post-processed using mean removal, length normalization and within-class covariance normalization [130]. Finally, for evaluation, probabilistic linear discriminant analysis (PLDA) [131] modelling was employed to compute the recognition scores. It should be noted that all of the aforementioned data sets consist of utterances spoken in a normal speech style. Most of the utterances were conversational telephone speech recordings.

### 6.2.2 PSM–GMM processing

In the experimental work presented in our first topic, we employ the PSM–GMM algorithm for normal-to-shouted SSC. A flowchart of the PSM–GMM algorithm appears in Figure 8. It is almost the same as the one presented in Section 5.2, but this one includes the features extracted from the source speech (here normal speech) that are used in this specific task to train the GMM. The settings for this experiment are presented next.

For the feature extraction phase and the PSM technique, we employ an all-pole mapping filter $H(z)$ of order $p = 12$, and we perform mel-warping with warping coefficient $\zeta = 0.31$. This value was shown in [132] to be suitable for mel-scale warping in case of narrow-band speech with sampling frequency of 8 kHz. Additionally, we use a filter bank of $M = 20$ channels. The mapping utilizes six-component, full-covariance GMMs.

At the training step, the GMM is trained using the expectation–maximization algorithm implemented in the GMMBAYES toolbox [133] with a data set of normal–shouted speech frame pairs. This EM implementation uses fuzzy $c$-means clustering

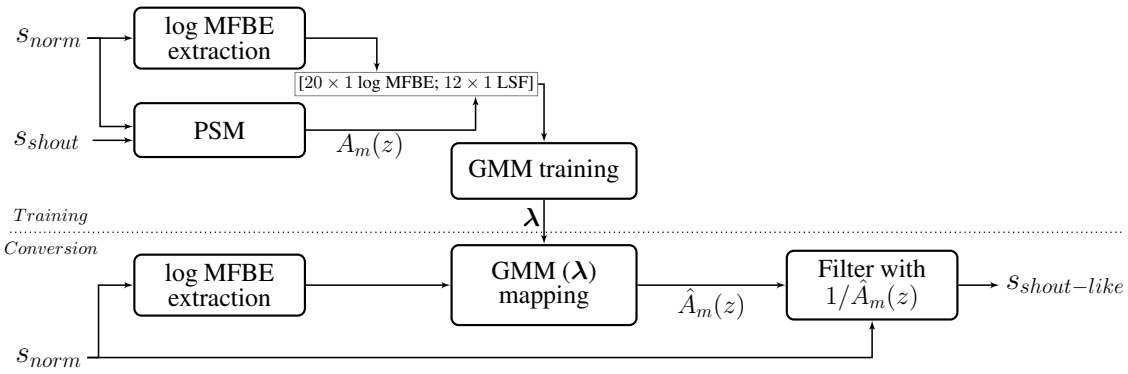


Figure 8: Flowchart of the PSM–GMM algorithm for converting frames of normal speech to shout-like speech. The output is the shout-like speech frame $s_{shout-like}$, which is obtained by filtering the normal speech frame $s_{norm}$ with the predicted mapping filter $\hat{H}(z) = 1/\hat{A}(z)$.

to initialize the component means. The component weights are set to have equal initial values, and the component covariance matrices are initialized to the diagonal vector of the full training data's covariance matrix [133]. Since the utterances of normal and shouted speech have different durations, the shouted speech frames are aligned to the normal ones using the DTW algorithm implemented in [134]. Although not completely accurate, this method is straightforward to employ for this kind of task. The feature vectors fed to the GMM for training are 32-dimensional, and they consist of the concatenation of two vectors:

1. A 20-dimensional log-MFBE vector of normal speech (for $\boldsymbol{x}^{(s)}$), and
2. A 12-dimensional vector of LSFs, which corresponds to the $A(z)$ filter (for $\boldsymbol{x}^{(t)}$).

Once the GMM is trained, PSM–GMM is applied to the normal speech frames at processing time to convert them into shouted speech. That is, normal speech frames are filtered with the predicted mapping, warped filter $\hat{H}_m(z) = 1/\hat{A}_m(z)$. However, the warped filter cannot be applied directly since it presents a delay-free loop. Thus, we obtain a usable version of the filter with the WarpTB toolbox [135]. It is important to note that prior to filtering, if the predicted filter $\hat{H}_m$ is unstable, its poles outside the unit circle must be replaced by the corresponding mirror image roots inside the unit circle, in order to resolve the instability problem.

## 6.3 Evaluation

We evaluate the normal-to-shouted speech conversion from our proposed SSC system in a text-independent speaker recognition task under vocal effort mismatch. The mismatch arises from a speaker recognition system (trained with normal speech, as mentioned in Section 6.2.1) in which speech utterances are shouted at test time but recorded in normal speech at enrollment. From now on, we will denote this

as the shouted vs. normal ($S - N$) scenario. Since this work deals with a speaker recognition task, its experiments are evaluated in terms of speaker identification rates.

For evaluation, we arrange the utterances as in the scheme presented in [118]: We pick half of the 24 utterances available for each speaker and use them at enrollment, and we use the remaining utterances at test time. We repeat this scheme of using 12 utterances at enrollment and the other 12 at test time by employing a circular rotation procedure. After completing the rotation, we have 12 sets of enrollment–test utterance pairs for each speaker. Since there are no cross-gender comparisons, we have a total of $12 \times 11 = 132$ comparisons for each gender.

In this work, we focus on the mismatch in vocal effort between test and enrollment utterances—that is, the shouted vs. normal scenario ($S - N$). Once PSM–GMM is applied to the enrollment utterances, we obtain the new scenario of shouted speech at test time and shouted-like speech at enrollment (shouted vs. shouted-like—that is, the $S - \hat{S}$ scenario). By comparing these two scenarios ($S - N$ vs. $S - \hat{S}$), we can evaluate how much speaker identification performance improves after applying PSM–GMM to the enrollment data.

In addition to evaluating the differences between $S - \hat{S}$ and $S - N$, we evaluate and compare the scenario of normal speech at both test time and enrollment (normal vs. normal, $N - N$) to the scenario of shouted-like speech at both test time and enrollment (shouted-like vs. shouted-like, $\hat{S} - \hat{S}$—that is, when both test and enrollment utterances are processed with PSM–GMM). The reduction in speaker identification rates between the $N - N$ and $\hat{S} - \hat{S}$ scenarios indicates how much speaker-dependent information is lost due to the PSM–GMM processing of the speech.

Finally, for the $S - \hat{S}$ scenario (where at enrollment, the level of normal speech is raised to that of shouted speech in terms of MFBEs by applying PSM–GMM), we introduce two evaluation conditions. These conditions are determined by the type and amount of data available to train the GMM used for mapping. They are as follows:

**SD**: A speaker-dependent (SD) GMM is trained and used during PSM–GMM processing for each trial of speaker recognition. Under this condition, only 12 utterances, all considered to be from enrollment, and their shouted-speech counterparts are available to train the GMM.

**GD**: A gender-dependent (GD) GMM is trained and used during PSM–GMM processing. That is, all the normal and shouted speech utterances for each gender are available to train the corresponding GD GMM.

In addition to the SD and GD conditions, the $S - \hat{S}$ scenario requires a condition for evaluating the efficiency of the PSM processing without including the GMM mapping. The condition we use for that is the following:

**Oracle**: We term this condition *oracle* because it assumes that the shouted speech version of each normal speech utterance at hand is available. Even though GMM mapping does not affect the oracle condition, other potential sources of

inaccuracy arise from the PSM processing technique. These include the use of the NNLS algorithm (Eq. 37), the assumption of piecewise linearity for the spectrum of the mapping filter, and the suboptimal alignment of frame pairs using DTW.

## 6.4 Results

The results from our experimental work on normal-to-shouted speech conversion (in terms of correct speaker identification rates) are shown in Table 1.

Table 1: Speaker identification rates (%) for males (M), for females (F), and on average (All). The table in (a) reflects the baseline (no processing) evaluation condition, while that in (b) reflects PSM–GMM processing evaluation conditions. The conditions SD and GD stand for speaker-dependent and gender-dependent, respectively, as defined at the end of Section 6.3. On the other hand, the different test speech vs. enrollment speech scenarios are denoted with letter pairs: the first letter corresponds to the test speech, and the second one to the enrollment speech; $\boldsymbol{N}$, $\boldsymbol{S}$, and $\hat{\boldsymbol{S}}$ stand for normal speech, shouted speech, and shouted-like speech, respectively.

(a) Baseline (no processing) condition.

| **Test-Enroll** | $\boldsymbol{N}$–$\boldsymbol{N}$ | | | $\boldsymbol{S}$–$\boldsymbol{N}$ | | |
|---|---|---|---|---|---|---|
| | **M** | **F** | **All** | **M** | **F** | **All** |
| Baseline | 95.5 | 96.2 | 95.8 | 62.1 | 23.5 | 42.8 |

(b) PSM–GMM processing conditions.

| **Test-Enroll** | $\hat{\boldsymbol{S}}$–$\hat{\boldsymbol{S}}$ | | | $\boldsymbol{S}$–$\hat{\boldsymbol{S}}$ | | |
|---|---|---|---|---|---|---|
| | **M** | **F** | **All** | **M** | **F** | **All** |
| Oracle | 93.2 | 93.2 | 93.2 | 88.6 | 65.2 | 76.9 |
| SD | 97.7 | 92.4 | 95.1 | 80.3 | 66.7 | 73.5 |
| GD | 93.9 | 96.2 | 95.1 | 46.2 | 37.1 | 41.7 |

When comparing the mismatch conditions of the non-processed and PSM–GMM-processed (that is, $S - N$ vs. $S - \hat{S}$) scenarios, one can see that the oracle and SD conditions provide a substantial improvement in speaker identification rate over the baseline condition for both genders. Examining the oracle condition of the $S - \hat{S}$ case, where no GMM mapping is applied, we see that the PSM technique by itself holds potential in converting the spectral properties of normal speech to those of shouted speech, since the identification rate for this condition is high (76.9% on average). Though this value is well below that of the non-mismatch ($N-N$) scenario (which has the average identification rate of 95.8%), the identification rate of the oracle condition is by a large margin better than that of the non-processed mismatch scenario (42.8% on average).

Once we introduce GMM mapping, we observe that for the SD condition, the speaker identification rate is lower than that of the oracle condition, but it is close enough (73.5% on average). This result implies that GMM mapping (Eqs. 11–13) and its implementation (Figure 8) are able to provide a fairly accurate mapping filter $\hat{H}(z)$ that, once applied to normal speech, can convert it successfully into shout-like speech. However, in the case of the GD condition (gender-dependent GMM modeling), the speaker identification rate does not improve over that of the baseline condition. The reason behind this may be the smoothing produced by the GMM mapping of a fixed model order over the data of all speakers.

Finally, when comparing the recognition rates between the $N$–$N$ and $\hat{S}$–$\hat{S}$ scenarios (that is, the non-mismatch scenario without or with PSM–GMM processing), the results show that PSM–GMM processing of normal speech does not cause a substantial loss of discriminant speaker information, since the recognition rates do not drop too much compared to the baseline case. Additionally, it must be noted that when looking at the mismatch conditions ($S-N$ and $S-\hat{S}$), the speaker identification rates are always lower for females than for males. This is most likely due to the effect of bias in the spectral envelopes caused by harmonics. That effect is stronger for high-pitched utterances, and females tend to have higher-pitched voices than males.

# 7 A vocoder-based parametric SSC system

This section presents the theory behind the SSC method proposed as part of the second topic of this thesis (normal-to-Lombard speech conversion). The SSC approach we employ is vocoder-based and parametric.

This SSC system utilizes supervised statistical mapping for automatic speech conversion. Thus, prior to performing the SSC, the system needs to be trained. Figure 9 shows the training and conversion stages of this SSC system. We present the system in a generalized form (with no specific source or target style) since it can be applied to conversions of different styles.

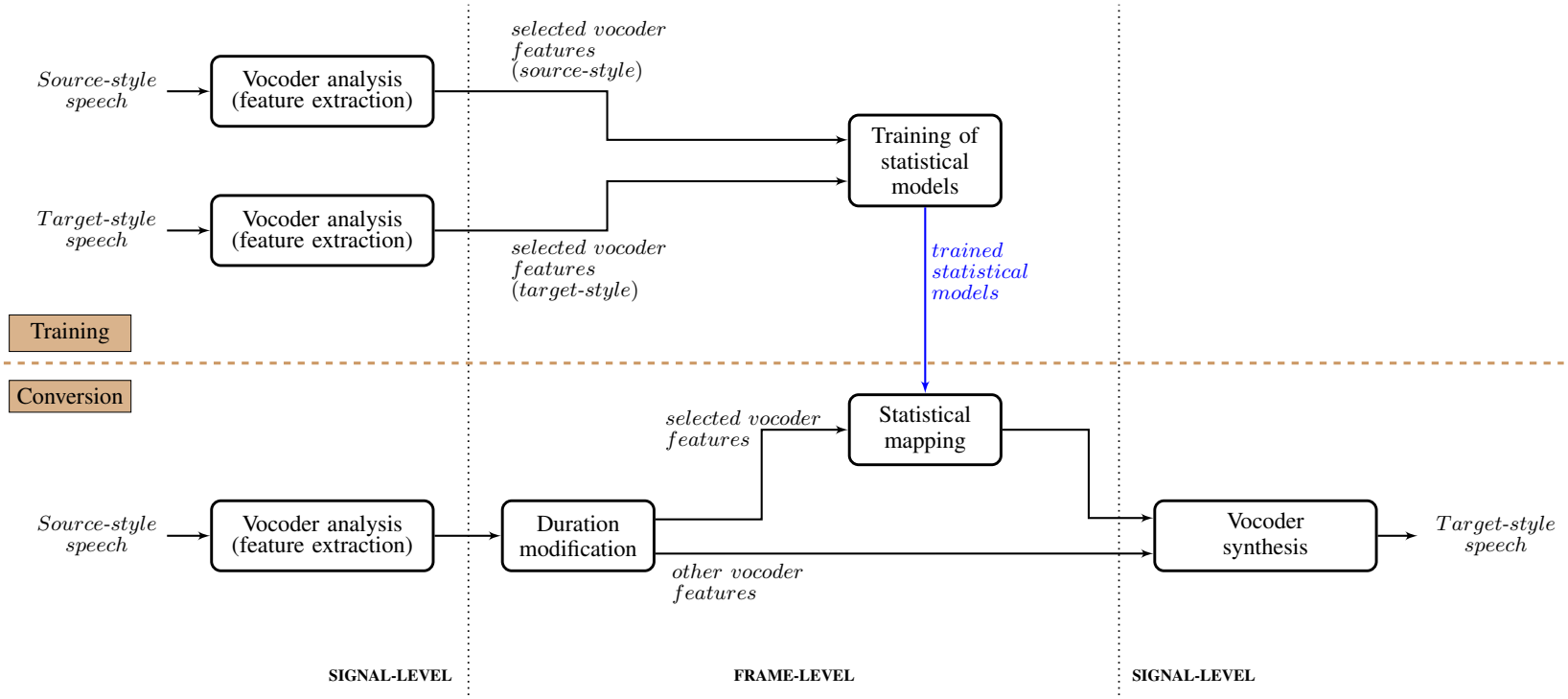


Figure 9: Training and source-to-target-style conversion of the vocoder-based parametric SSC system proposed. Prior to training the statistical models (one per selected vocoder feature), the source and target speech frames are aligned to obtain the source–target frame pairs.

For training, we first extract the speech features of interest using the vocoder selected for the system. We hereafter refer to such features as vocoder features. Next, given that mapping is done at frame level and that sometimes utterances have different durations in different speaking styles (for example, Lombard speech usually has longer duration than the corresponding normal speech), we align the source and target frames using using DTW [134]. Finally, we train the statistical models used for mapping (one model per selected feature) to learn how to transform each of the selected vocoder features of the source-style speech into the corresponding features of the target-style speech. We use the aligned frame pairs of source–target speech as training data after discarding the aligned source–target frames that are classified in the opposite voiced/unvoiced categories.

We use the vocoder-based parametric SSC system described here specifically to convert normal speech into Lombard speech. For that reason, of all the available vocoder features, we select for mapping only those pertaining to the speech attributes that affect the generation of Lombard speech. Those attributes are 1) spectral tilt, 2) $f_0$, 3) energy, and 4) duration. However, apart from spectral tilt, vocal tract modifications are also key to Lombard speech. Specifically, these are the shifting of formant frequencies and the narrowing of their bandwidths. That said, to maintain simplicity in our system (which represents a novel method) we do not consider the vocal tract attribute here. We map the vocoder features representing spectral tilt, $f_0$, and energy using one statistical model per feature. The spectral tilt and $f_0$ features are mapped only for voiced frames, while the energy feature is mapped for both voiced and (active, non-silent) unvoiced frames. The voiced/unvoiced frame decision is based on the $f_0$ value, while the silent frames are identified using both $f_0$ and an energy threshold criterion.

We modify the duration attribute by scaling the duration of the voiced and unvoiced regions separately. Then, we apply this modification to all our vocoder features. For those features that we map, we perform the duration modification first. To compute the scaling value, we first align the frames of source style and target style for the same utterance, using DTW. Then, we obtain the scaling value as the mean ratio of the locations of the frames of the source style to those of the corresponding frames in the target style. Outliers in the ratios likely originate from inaccuracies in the frame alignment process, so we exclude them. Finally, prior to synthesis, we smooth the trajectories of the mapped features (except $f_0$) using a moving average filter. This reduces distortions in the converted utterances.

The method at hand is implemented using two vocoders as the foundation of the conversion system to compare the two cases. Meanwhile, the statistical mapping technique we employ is the same in both cases. The following sections elaborate on two aspects of our method:

1. The vocoders we compare: glottal vocoder and STRAIGHT. Both of these are often used in SPSS.

2. The statistical mapping model employed with both vocoders: BGMM.

## 7.1 Vocoder framework

Our proposed method includes a glottal vocoder as the framework for the analysis and synthesis steps. However, given our interest in evaluating the impact of the vocoder on SSC performance, we also study another vocoder framework: the widely used STRAIGHT vocoder. Consequently, we adapt the SSC system employed here to each vocoder case.

### 7.1.1 Glottal vocoder framework

The first vocoder used is a state-of-the-art variant of the glottal vocoder presented in [28]. Our implementation of the glottal vocoder produces the glottal flow pulses employed during synthesis from a DNN model [55]. In the current SSC task, we have access to the original speech signal and, after applying GIF to the voiced frames, we gain access to the original estimated glottal flow signal. This contrasts with standard text-to-speech tasks, since the vocoding approach in [55] does not rely on waveform modelling, instead employing the estimated glottal waveforms directly. This procedure differs from standard vocoding insofar as non-parametric information is used during the synthesis step. Therefore, our approach can be understood to be closely related to LP-residual PSOLA [136].

During the analysis step, the glottal vocoder extracts the following features: 1) log energy, 2) HNR, 3) $f_0$, 4) vocal tract LSFs ($LSF_{VT}$), and 5) glottal source LSFs ($LSF_{glott}$). The glottal vocoder features selected for normal-to-Lombard conversion are: the $LSF_{glott}$ (for spectral tilt), $f_0$, and energy vocoder features. Each of these features is mapped separately via a mapping model (specifically, a BGMM).

It should be noted that decomposing speech into glottal excitation and vocal tract is difficult and might not always succeed perfectly for all speech frames. Therefore, the residue of the spectral tilt present in the vocal tract is measured and applied to the glottal-source spectral tilt. This is done by:

1. Computing first-order LP of the vocal tract model (an all-pole filter denoted by $H_{VT1}(z)$),
2. Filtering the LP coefficients corresponding to $LSF_{glott}$ with $H_{VT1}(z)$ and obtaining the new feature $LSF'_{glott}$,
3. Modifying the modified feature $LSF'_{glott}$ in duration (like the rest of features),
4. Mapping each of the selected vocoder features with a BGMM,
5. And, finally, after mapping, filtering the predicted spectral tilt of the glottal source's LP model for Lombard speech with $1/H_{VT1}(z) = A_{VT1}(z)$ (where $A_{VT1}$ has also been modified in time) to compensate for the tilt included in the non-modified vocal tract spectrum feature used in synthesis.

Additionally, it should be noted that included within the glottal vocoder construction is a smoothing operation on the feature trajectories (except $f_0$) during

synthesis. This can prove helpful when synthesizing converted speech from the predicted features, as it reduces artifacts. Flowcharts representing the training and conversion stages of the normal-to-Lombard speech conversion task using the glottal vocoder are presented in Figures 10 and 11, respectively.

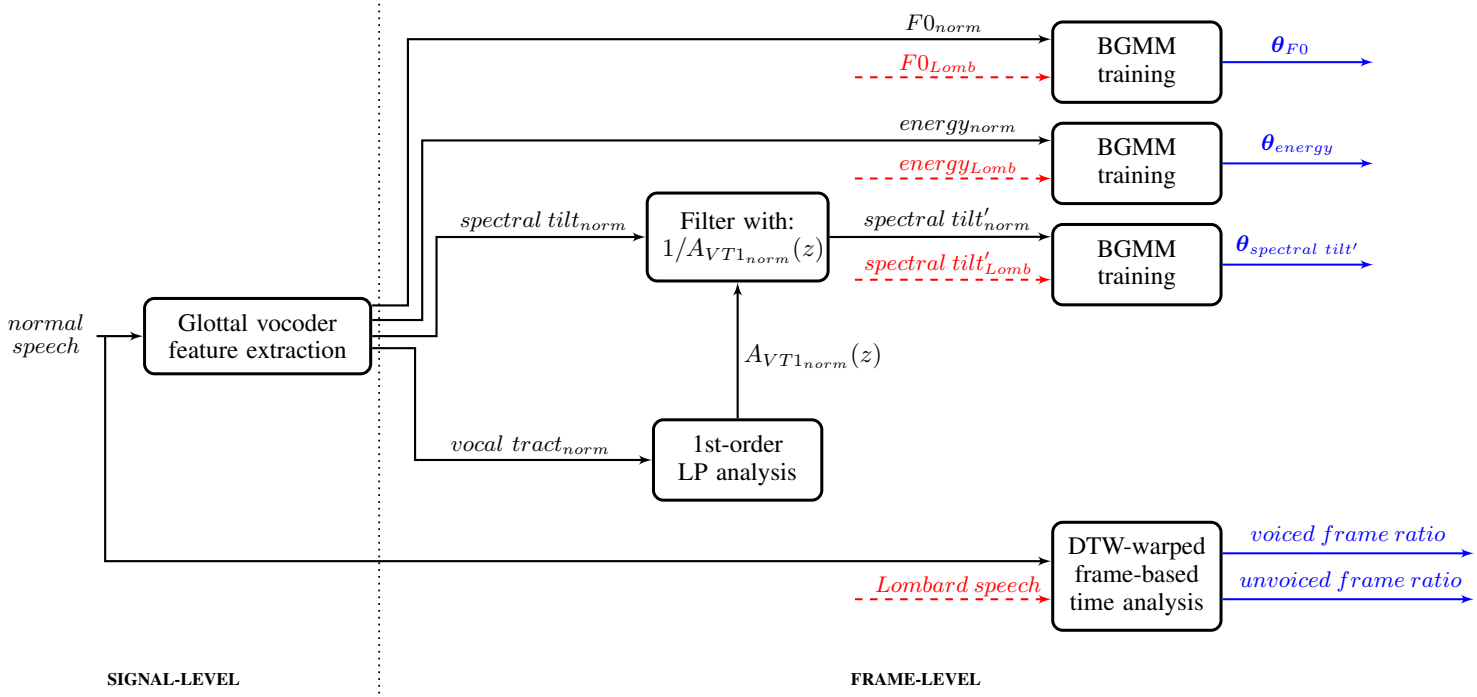


Figure 10: Block diagram of the training stage of the glottal-vocoder-based, normal-to-Lombard speech conversion system. Features $LSF_{glott}$ and $LSF_{VT}$ are here called *spectral tilt* and *vocal tract* for the system to be easily understood at first glance. $H_{VT1}(z) = 1/A_{VT1}(z)$ is the first-order all-pole filter obtained from the vocal tract model. This is understood to be the spectral tilt residue in the vocal tract. The red dashed arrows indicate features obtained from Lombard speech in a similar manner to normal speech, since for BGMM training, DTW-aligned pair frames from normal speech (source style) and Lombard speech (target style) are needed. In blue are the parameters used at conversion time for BGMM mapping as well as the voiced and unvoiced ratios utilized duration modification. $\boldsymbol{\theta}$ denotes the BGMM parameters.

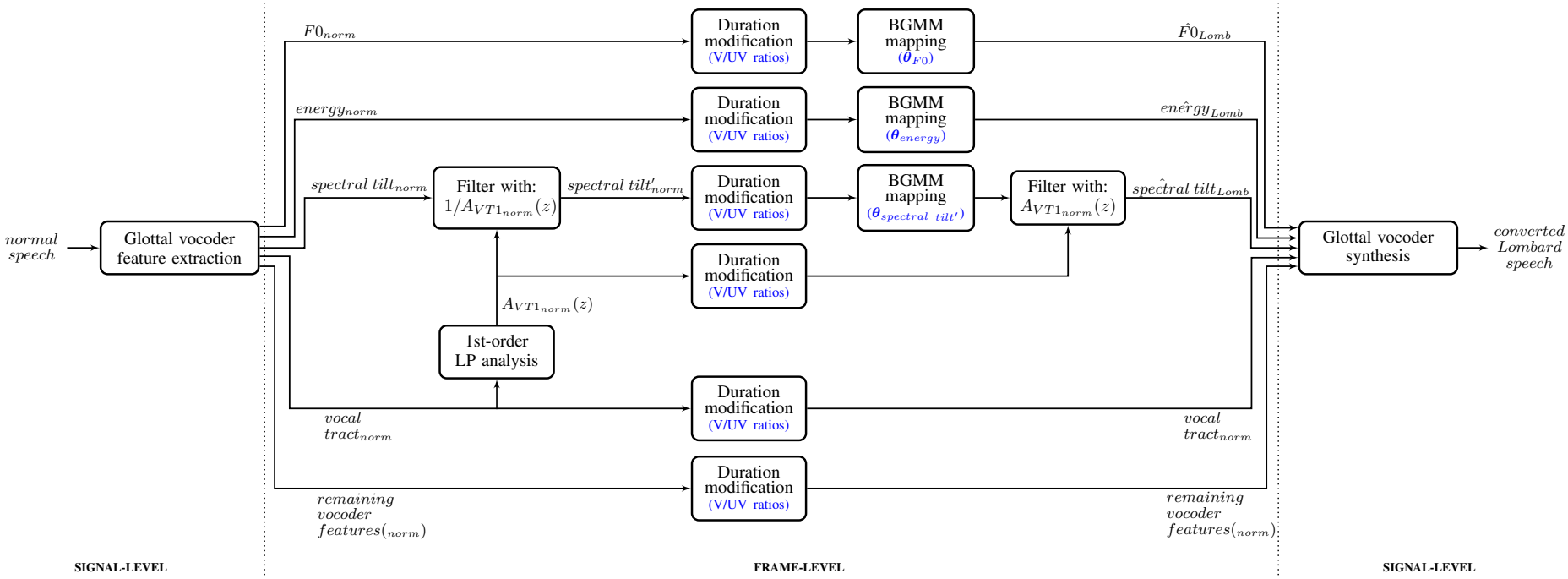


Figure 11: Block diagram of the conversion stage of the glottal-vocoder-based, normal-to-Lombard speech conversion system. Features $LSF_{glott}$ and $LSF_{VT}$ are here called *spectral tilt* and *vocal tract*. $H_{VT1}(z) = 1/A_{VT1}(z)$ is the first-order all-pole filter obtained from the vocal tract model. In blue are the BGMM parameters ($\boldsymbol{\theta}$) used for mapping as well as the ratios of voiced (V) and unvoiced (UV) frames used for duration modification. Both the BGMM parameters and the ratios were obtained at the training step.

### 7.1.2 The STRAIGHT vocoder framework

The STRAIGHT vocoder extracts the following features to represent speech: 1) the ABEs, which represent the aperiodicity spectrum, 2) $f_0$, and 3) the spectral envelope, which is represented using an MGC. More details on this vocoder are presented in Section 2.3.2.

Since our goal is to make a meaningful comparison with the glottal vocoder case, we select the same speech attributes for mapping in the STRAIGHT case (the energy, $f_0$, and spectral tilt). As feature $f_0$ is an element of the set of STRAIGHT features, its modification is straightforward. We change the spectral tilt by mapping only the first two mel cepstrum coefficients ($c_1$ and $c_2$) of the MGC feature, leaving the remaining coefficients unchanged (as in [137]). Because the STRAIGHT vocoder does not include an explicit energy feature, we modify the energy by adjusting the final synthesized speech signal. To achieve this, in the training stage we compute log-energies of the frames of Lombard and normal speech and use them to train a separate BGMM. Then, at conversion time, we compute frame-level log-energies from the synthesized samples, and we scale the non-silent speech frames based on the Lombard log-energies predicted by the trained BGMM.

Prior to synthesis, to reduce distortions on the converted samples, we smooth the trajectories of the mapped features (except $f_0$) using a moving average. We then obtain the converted Lombard speech using the overlap–add method on the Hanning-windowed scaled speech frames. Sometimes, the global level of a given converted Lombard utterance in each of the STRAIGHT and glottal vocoder cases differs slightly (around 0.5 dB). Thus, to make sure that the comparison is fair, we adjust the global levels of the converted speech samples in the STRAIGHT case to be the same as those in the glottal vocoder case.

A flowchart showing the training steps in the STRAIGHT case is presented in Figure 12. The conversion stage for the STRAIGHT-based system is shown in Figure 13.

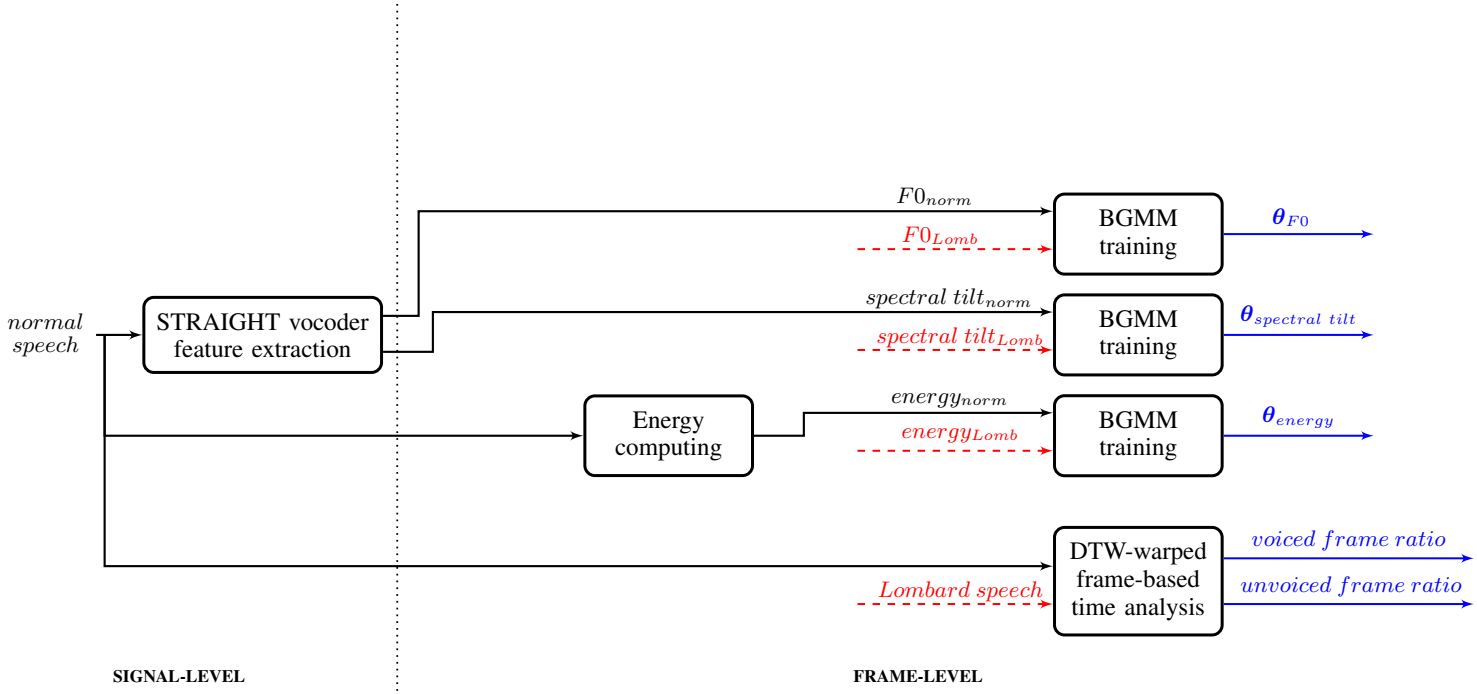


Figure 12: Block diagram of the training stage in the case of the STRAIGHT-based system proposed for normal-to-Lombard-speech conversion. Here, *spectral tilt* represents the coefficients $c_1$ and $c_2$ of the MGC vocoder feature. The red dashed arrows indicate features obtained from Lombard speech in a similar manner to that of normal speech, since for BGMM training, DTW-aligned pair frames from normal speech (source style) and Lombard speech (target style) are needed. Additionally, the parameters used at conversion time for BGMM mapping are marked in blue, as are the voiced/unvoiced ratios employed for duration modification. $\boldsymbol{\theta}$ denotes the BGMM parameters. It should be noted that since the time analysis block is not dependent on the vocoder, this step concludes with the same ratios obtained for the glottal vocoder case. Thus, the step is not needed if the ratios are available beforehand, as they are here.

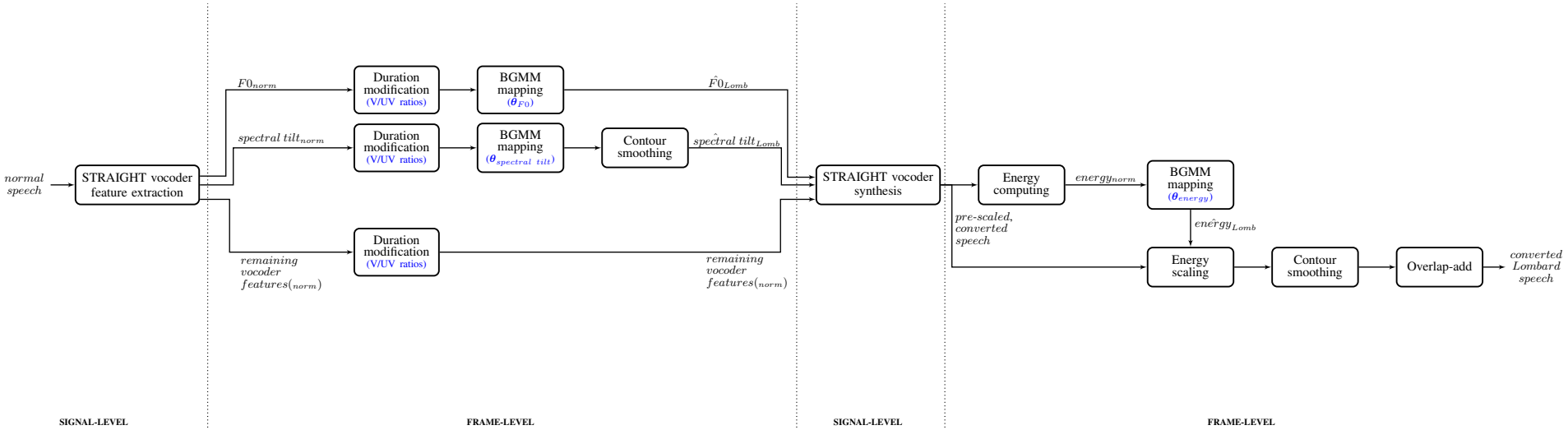


Figure 13: Block diagram of the conversion stage in the case of the STRAIGHT-based system proposed for normal-to-Lombard-speech conversion. Here, *spectral tilt* represents the coefficients $c_1$ and $c_2$ of the MGC vocoder feature.

## 7.2 Statistical mapping BGMMs

To further investigate the second topic of this thesis (conversion of normal speech to Lombard speech), we propose an SSC system that employs a statistical model for mapping source speech (normal speech) to target speech (Lombard speech) along a vocoder for analysis and synthesis. Specifically, we employ one statistical model per selected vocoder feature and map each feature in source style ($\mathbf{y}^{(s)}$) to its counterpart in target style ($\mathbf{y}^{(t)}$). As in the system proposed in Section 5, the mapping procedure for this SSC system is supervised. This means that parallel speech data from the source and target speaking styles ($\mathbf{x}^{(s)}$ and $\mathbf{x}^{(t)}$, respectively) are needed to train the mapping model. As stated previously, our goal is to convert normal speech into Lombard speech. Unfortunately, the availability of data on Lombard speech (and other speaking styles) are rather limited since the procedure for recording the speech signals is quite elaborated when having to record source–target parallel data. Additionally, such recording sessions may be harmful to speakers' health, since in the case of high-vocal-effort speaking styles, speakers must raise their voices for relatively long periods of time.

To cope with the restriction of speech data in the target style (Lombard speech), we decided to use BGMM as a statistical mapping model. This model is an extension of the GMM employed in the SSC system from Section 5. Compared to GMMs, BGMMs can cope better with limited data and are less influenced by the over-fitting effect. Given the intractability of the posterior distribution of the BGMM parameters (that is, the density function needed to predict target features during the conversion stage), we employ variational inference to obtain an approximation of the posterior. A detailed explanation of BGMMs and variational inference is presented in Section 3.1.2.

# 8 Experimental work (topic II): Normal-to-Lombard-speech, vocoder-based parametric SSC using Bayesian GMMs

The experimental work for our second topic utilizes the SSC system proposed in Section 7 to carry out conversion of normal speech into Lombard speech. This conversion system is vocoder-based and parametric. The goal of this experimental work is to determine whether the proposed system is able to convert the speech signal. We additionally study the impact on conversion of the vocoder used in the proposed system. As mentioned in the previous section, for this reason, we utilize two different vocoders. The first is a glottal vocoder that splits speech into a vocal tract filter and glottal flow excitation [28]. We compare that vocoder to the widely known STRAIGHT vocoder [40]. Our goal is to evaluate the speech quality of the system when employing each vocoder. Given that the purpose of the glottal vocoder is to parametrize the two main parts of the production of natural speech (vocal tract and glottal flow), we hypothesized that the system based on the glottal vocoder would perform better and produce more natural converted Lombard speech.

## 8.1 Data

Our data set for this experiment consists of recordings of 10 Finnish speakers: four females and six males [94]. Each recording consists of a speaker reading a text of 90 words over the course of approximately one minute. Each recording is produced using two different levels of vocal effort: 1) a normal vocal effort, which produces normal speech, and 2) an increased vocal effort in a noisy situation, which results in Lombard speech. For each speaker and level of vocal effort, we have recordings of 11 utterances. We downsampled the recordings from 48 kHz to 16 kHz for our experiment.

## 8.2 Experimental setup

First, to extract the features, we used analysis frames of 25 ms with a frame shift of 5 ms. In the glottal vocoder case, the $LSF_{glott}$ feature was 10-dimensional, the $LSF_{VT}$ feature was 30-dimensional, and the HNR feature had 5 frequency channels. Both the fundamental frequency feature ($f_0$) and the glottal closure instants used in the QCP method were computed using the REAPER tool [138] under its default settings (except for the frame shifting specified at 5 ms). In the case of the STRAIGHT vocoder, the extracted features were the following: the first 40 MGC coefficients, excluding the log-energy coefficient $c_0$, and 21 aperiodicity energy bands. As described in Section 7.1.2, of those 40 MGC coefficients, only $c_1$ and $c_2$ were mapped for conversion, while the other coefficients remained unchanged [137]. For this vocoder, $f_0$ was also extracted using the REAPER tool [138] under its default settings.

The utterance durations were modified by scaling the durations of the voiced and unvoiced regions separately, using frame-based interpolation of all the vocoder

features prior to BGMM-based mapping. Cubic spline interpolation was used on all features of the two vocoders. As described in Section 7, the scaling values for the unvoiced and voiced regions were each computed as the mean ratio of the locations of the DTW-based aligned frames of the corresponding source and and target speech samples. The scale values obtained from the training data were 1.08 and 0.88 for the voiced and unvoiced regions, respectively. Thus, the voiced and unvoiced regions were respectively stretched and compressed after the interpolation. The scale values we found are in line with those of earlier work [60].

To map the selected vocoder features of normal speech to their counterparts in Lombard speech, we utilized BGMMs with $J = 100$ components. Due to the small size of our data set, the BGMMs were trained for each vocoder feature and each speaker using a training set consisting of the utterances of all remaining speakers (both males and females) in the original set. Because BGMM training requires parallel data, we used feature pairs of normal and Lombard speech frames to train each BGMM and map the corresponding feature. To find the frame pairs, we aligned the normal and Lombard speech frames using DTW [134] due to differences in the durations of corresponding utterances in each speaking style.

One advantage of BGMMs is that they do not suffer from the over-fitting effect even in the presence of a large number of Gaussian components $J$ (and therefore a large number of model parameters). That is, BGMMs do not increase test error as complexity increases, meaning that the value of $J$ is not too critical so long as it is sufficiently large. For the sake of simplicity, we used BGMMs of $J$=100 components for all vocoder features. A 10-fold cross-validation check showed that using a larger number of components did not yield significant improvements in terms of the root-mean-square (RMS) error.

We modelled the component means and precisions of the BGMMs with prior distribution $\mathcal{NW}(\mu_0, \beta_0, W_0, \nu_0)$, setting the prior parameters similarly to those recommended in [139]. We let $\mu_0$ and $W_0$ equal the data set mean and precision, respectively, and we set $\beta_0 = 1$ and $\nu_0 = D + 2$. We set the concentration parameter $\alpha_0$ equal to an all-ones vector.

After completing the mapping, we obtained the converted Lombard samples via synthesis with the corresponding vocoder.

## 8.3 Evaluation

We evaluated our proposed conversion system and compared the two versions based in different vocoders with the help of two listening tests. These tests evaluated the quality of the converted speech samples in terms of 1) similarity of the converted Lombard speech to natural Lombard speech and 2) naturalness of the converted speech samples. The listening tests were carried out under a modified version of the BeaqleJS evaluation framework [140]. BeaqleJS (a browser-based evaluation of audio quality and a comparative listening environment) provides a framework based on open web standards such as Javascript and HTML5, which enables the creation of browser-based listening tests.

The first listening test was a similarity test to evaluate the perceptual proxim-

ity of the converted Lombard speech (vocoded using either the glottal vocoder or STRAIGHT) to naturally produced Lombard speech. In this evaluation, listeners were asked to rate how much each converted sample resembled a naturally produced Lombard sample. A continuous scale from 1 to 5 was used for rating, with the scale numbers describing the resemblance between samples as follows: 1 – none, 2 – little, 3 – moderate, 4 – much, 5 – very much.

For the evaluation of each test case, listeners were provided with a piece of non-converted reference speech produced by vocoding the utterance of the test case at hand in a normal speaking style. During the evaluation, listeners were allowed to hear each sample as many times as they wished before continuing to the next test case. This task utilized 16 randomly selected utterances from the data set. Of these utterances, half were produced by four females speakers, and half were produced by four male speakers. Each speaker contributed two utterances. Since listeners rated the proposed conversion system for each of the two vocoders, each listener rated 32 test cases in total. These cases were presented to listeners in random order. Prior to the evaluation, the listeners participated in a familiarization session, where they were introduced to how Lombard speech sounds. In that session, the listeners heard a few examples of normal vs. Lombard speech. The samples employed for this purpose were not reused for the evaluation. Also during the familiarization session, listeners were asked to adjust the volume in their headphones to a loud yet comfortable level and were instructed to maintain that level for the test.

The second listening test was a preference test on naturalness. For that, a pairwise comparison was done to evaluate the naturalness of converted Lombard speech samples produced by the glottal vocoder and by STRAIGHT. During each test case, listeners were presented with two versions of the same utterance. One was a speech sample converted using the glottal vocoder, and the other was a sample converted using STRAIGHT. The versions were presented in random order, meaning that it could never be assumed that a given sample corresponded to one vocoder or the other. Listeners were asked to choose which sample sounded more natural, and they were also allowed to indicate if they had no preference between the two. As in the first test, listeners were permitted to hear each sample as many times as they wished before moving on. During this test, each listener evaluated 24 cases. These 24 utterances were selected randomly from the data set. The utterances belonged to four female and four male speakers, meaning that each speaker contributed three utterances.

## 8.4 Results

The results of the similarity test are summarized in boxplot form in Figure 14. Each box's central line (i.e. the notch in the middle) represents the median, while the lower and upper edges of each box represent the first and third quartiles (Q1 and Q3), respectively. The notch area, computed using Gaussian-based asymptotic approximation [141], indicates the 95% confidence interval of the median. The whiskers extending from each box delineate the most extreme points in the data set of our results. The red, diamond-shaped points indicate outliers in the data.

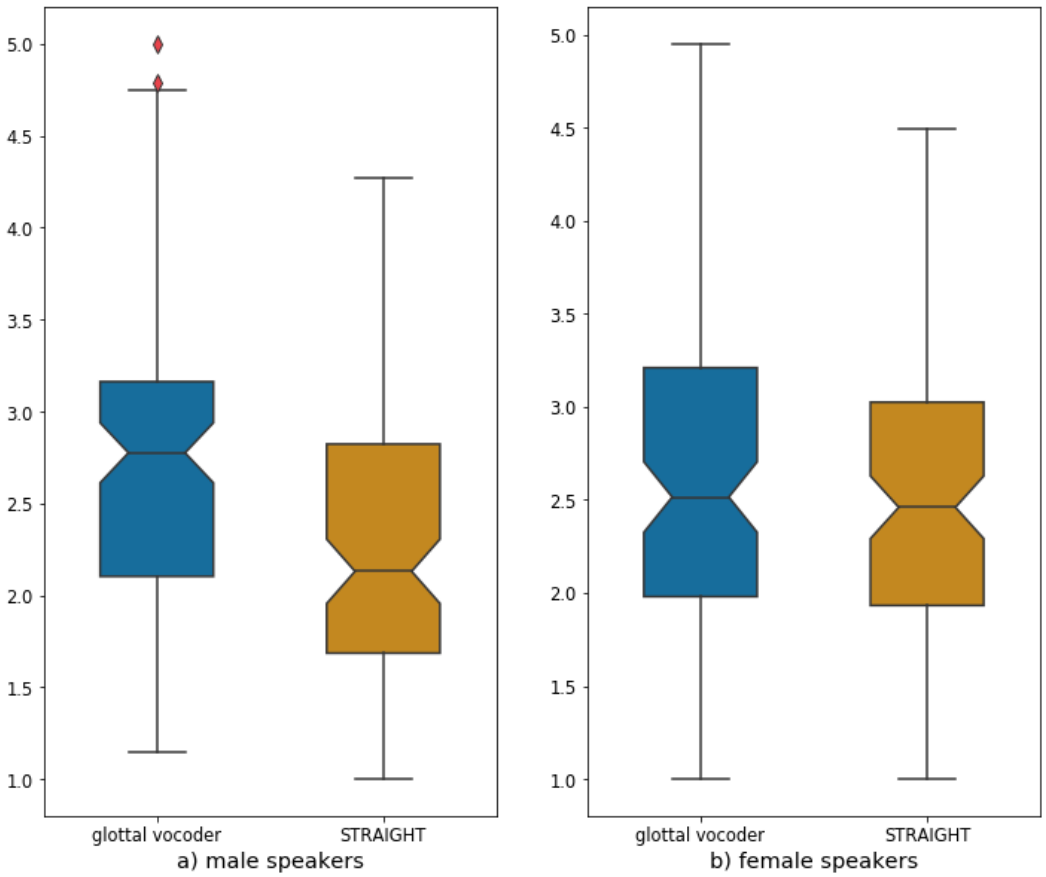


Figure 14: Results of the similarity test presented in boxplot form. In this test, listening subjects evaluated the resemblance of converted speech samples to naturally produced Lombard speech samples. In this visual representation, the median is indicated by the middle line in each box. The notch around that line represents the 95% confidence interval around the median. Outliers in the data appear as small, red, diamond-shaped points. The rating scale for resemblance between speech samples was: 1 – none, 2 – little, 3 – moderate, 4 – much, 5 – very much.

The results in Figure 14 indicate that both the glottal-vocoder-based and STRAIGHT-based conversion systems are able to transform normal speech into Lombard speech. However, discrepancies arise when considering the results genderwise. According to the book *Graphical methods for data analysis* [142], if the notched areas (computed as in [141]) of two boxes in a boxplot do not overlap, there is a strong chance that a similar difference in median levels would occur in other data sets collected under similar conditions. In other words, lack of overlap in the notched areas can be interpreted as evidence that the medians differ. In our experiment, this overlap is absent in the case of male speakers, where the glottal vocoder has a larger median than STRAIGHT. However, for female speakers, the notched areas overlap, indicating that the median values are almost identical for the two vocoders.

The results of the naturalness preference test are presented in Table 2. These results show the percentage of listeners that preferred the STRAIGHT- vs. glottal-

vocoder-based conversion samples in the separate cases of male and female speakers. Listeners strongly preferred the samples corresponding to the glottal vocoder (98.61% for male speakers, 97.92% for female speakers) over those corresponding to STRAIGHT (0.69% for male speakers, 0% for female speakers). In a small percentage of test cases (0.69% for male speakers, 2.08% for female speakers), listeners expressed no preference for either sample.

Table 2: Results of the pairwise comparison test of listener preference on the naturalness of converted speech samples. "Male" and "Female" correspond to the subsets of converted speech samples uttered by male and female speakers, respectively. The results are reported as percentages [%].

| | **Glottal vocoder** | **STRAIGHT** | **No preference** |
|---|---|---|---|
| Male | 98.61 | 0.69 | 0.69 |
| Female | 97.92 | 0.00 | 2.08 |

# 9 Discussion and conclusions

In this section, we present discussion points and conclusions on the field of SSC in general, and on the topic of conversion of normal speech to high-vocal-effort speech in particular. Our argumentation is based on the results from our experimental work and also on the results from other research works in the same topic. We organized the section as follows. First, Section 9.1 discusses the issue of terminology standardization in the field of SSC. Second, in Section 9.2, we draw conclusions about the topic of conversion of normal to high-vocal-effort speech. Third, Section 9.3 discusses and compares the two approaches commonly employed in SSC: direct transformation vs. vocoder-based parametric. Last, Section 9.4 focuses on the mapping techniques used in SSC.

## 9.1 Discussion of the definition of SSC

This thesis included a literature review on SSC, focusing on the conversion of normal speech into high-vocal-effort speech. From the studies cited here (and, generally, from other SSC studies), it can be seen that there is no accepted standard terminology related to SSC in published research. This may make referencing across works inconsistent and thus hamper progress in the field. After all, the absence of predictable, uniform nomenclature makes it difficult to do a comprehensive review of past studies. As an example of this problem, we note that the experimental works presented in this thesis [22, 27] and some others [76, 91, 95] favor the term SSC. There also exist works focused on high-vocal-effort SSC that refer to the task as *voice quality modification* (e.g. [10]). *Voice quality*, as defined by Laver in [143], is "the characteristic auditory colouring of an individual speaker's voice". By modifying their voice quality, speakers can express changes in intention, emotion, and attitude. Thus, *voice quality* and *speaking style* can be understood to be separate denominations for the same idea. In turn, both SSC and *voice quality modification* seem to be reasonable terms. Still other works dealing with SSC denote the task as *voice conversion* or *transformation* (e.g. [73]). VC has been commonly understood in the field of speech technology to mean the transformation of one speaker's identity to another one. Using such terminology for SSC tasks can technically be considered correct if the voice is considered to define speaker identity as well as speech attributes intrinsic to the speaking style. However, this usage is nevertheless confusing, as speaker identity is a key element of the VC concept in speech technology. On the whole, this thesis' author believes that the use of standardized terminology in the field of SSC (or voice quality modification) would improve the pace of progress in this area of speech technology.

## 9.2 Conversion of normal speech to high-vocal-effort speech

This thesis presented original research on two types of SSC of normal speech to high-vocal-effort speech. In the case of normal-to-shouted speech conversion, we proposed a direct transformation SSC approach and evaluated it via a task of speaker recogni-

tion under vocal effort mismatch. In case of normal-to-Lombard speech conversion, we proposed a vocoder-based parametric SSC system and evaluated it using human listeners.

In the first case, we proposed an SSC system to be applied to a speaker recognizer with vocal effort mismatch. The mismatch consisted of a speaker recognition system with enrollment utterances from normal speech and test utterances from shouted speech. The mismatch caused a decrease in speaker identification rates. To solve that problem, we mapped enrollment utterances to shouted speech (in terms of MFBE features) using our proposed system, PSM–GMM. Prior to performing this study, we hypothesized that using PSM–GMM for SSC would result in higher recognition rates than those obtained without SSC. We evaluated this experimental work by its resulting speaker identification rates, which showed that when the statistical mapping model was trained with speaker-dependent data, SSC substantially improved speaker identification rates.

Another goal in the study of our first case was to evaluate whether the SSC from PSM–GMM could be performed without degrading speaker identity. To that end, we compared the results of the $N$–$N$ scenario (normal speech in both enrollment and test utterances) to those of the $\hat{S}$–$\hat{S}$ scenario (converted shouted-like speech in both enrollment and test utterances). The $\hat{S}$–$\hat{S}$ results did not decrease much in comparison with those of $N$–$N$. Thus, discriminating information on speaker identity does not appear to be lost due to the PSM–GMM conversion.

The second type of speaking style conversion treated by this thesis was normal-to-Lombard speech SSC. The focus of our study was to determine whether a perceptible conversion could be achieved without sacrificing the quality of the speech signal. To evaluate this, we utilized listening tests with human participants. These tests quantified the amount of Lombard effect that listeners were able to perceive in converted speech samples, and they also evaluated the samples' subjective naturalness. The results indicated that the proposed system was able to convert normal speech to Lombard-like speech.

While this system was able to convert normal speech to Lombard-like speech, the level of Lombardness achieved in both vocoder cases was far from high. The SSC proposed modified the energy, duration, $f_0$, and spectral tilt of the source speech signal to produce the conversion. However, our results clearly indicate that applying these modifications alone does not suffice to produce clear Lombard-like speech. This conclusion is in line with previous studies on normal-to-Lombard speech conversion [72, 73]. In their study [72], Huang et al. evaluated the conversion system by testing converted speech samples' similarity to Lombard speech. They performed the conversion in two ways: 1) by using only individual modifications of selected speech features ($f_0$, duration, and spectral envelope) and 2) by combining all of these modifications. Their results showed that the individual modifications did not manage to convert speech signals into Lombard-like speech, while the combined modifications did. Work in [73] compared the levels of conversion achieved by modifying selected features (formants, $f_0$, and energy spectral distribution) individually and jointly. Those results showed that the best conversion was achieved via joint modification of the selected features. It should be noted that vocal tract spectral modifications are

also key in Lombard speech, but we did not include them for the sake of simplicity. Bearing all of this in mind, our results suggest that, in future work, the vocal tract attribute should be included in the modifications of the SSC system to improve the conversion level.

Section 4 presented a literature review on SSC of normal speech to high-vocal-effort speech. Regarding the novelty of the two experimental works presented in this thesis, it should be noted that at the time these two works ([22, 27]) were published, research on SSC of normal speech to high-vocal-effort speech was scarce [11, 9, 72, 10]. Furthermore, the aforementioned studies were mostly limited to evaluating SSC of single words [72], logatomes [10], and a few example sentences [11]. In contrast, the two experiments presented in this thesis focus on the conversion of speech utterances. The only previous study to make use of utterances was [9], but that study only utilized synthetic speech. The work presented in [22], to the best of our knowledge, had the novelty of being the first study on high-vocal-effort-focused SSC that employed the proposed conversion method in a task of speaker recognition under vocal effort mismatch.

## 9.3 SSC approaches: direct transformation vs. vocoder-based parametric

With the system we utilized in our second study (a vocoder-based parametric approach for normal-to-Lombard speech conversion), one of our aims was to investigate the impact of the vocoder used on the SSC method. First, we proposed a glottal vocoder, which decomposes speech into glottal source and vocal tract components. For comparison, we also used STRAIGHT, which is commonly employed in vocoder-based parametric SSC systems. The Lombardness levels for both vocoders were similar for females, but for males, the glottal vocoder yielded a higher value. In terms of naturalness, test subjects tended to prefer the glottal vocoder over STRAIGHT. This result could be explained by the fact that the STRAIGHT-converted speech samples were characterized by artifacts like buzzing that may be more disruptive to the human ear than the artifacts in the glottal-vocoder-converted samples. Our results showed that the vocoder plays a key role in SSC, and this is supported by the higher level of Lombardness obtained for males with the glottal-vocoder-based parametric SSC system. That same vocoder also produced more natural-sounding speech samples according to evaluation by human listeners. In other words, although we might have expected to obtain less natural converted samples from the glottal vocoder in the case of male speakers (given that the levels of Lombardness are higher for the glottal vocoder than for STRAIGHT), this did not occur. Thus, the trade-off between the amount of conversion achieved and the quality of the converted samples proved to have less of an impact in the case of the glottal vocoder than in the case of STRAIGHT. We thus conclude that the choice of vocoder significantly impacts the converted samples.

The influence of the choice of vocoder within a vocoder-based parametric SSC system has also been studied in [76]. For the study of the second topic, Seshadri et al. used three vocoders (a glottal vocoder, STRAIGHT and PML) and compared

them in a SSC framework for a normal-to-Lombard speech conversion task. The results showed that using PML achieved a higher level of Lombardness in converted speech than did using the other vocoders. The trade-off between quality and level of conversion has also been treated by past studies [10, 9, 95]. In the case of [10], in which conversions between soft, normal, and loud vocal efforts were performed, results showed a correlation between the degree of conversion achieved and the degree of degradation in the quality of the signal. For larger degrees of conversion (such as from loud to soft), test subjects rated the speech quality lower than in the case of smaller degrees of conversion (such as from loud to normal). Additionally, the study in [9], which performed conversion of normal speech to Lombard speech for intelligibility improvement, argued that the lack of intelligibility in one of the cases under study was a consequence of the aforementioned trade-off: The degradation in speech quality resulting from the conversion negated the improved intelligibility. Lastly, in [95], the authors reasoned that the lack of intelligibility improvement in some of their results may have been due to the operations performed during conversion (vocoding and mapping). Thus, conversion operations seem to impact not only speech quality but also intelligibility improvement (in the case of having Lombard speech as the target style).

As demonstrated in this thesis, there are two approaches to tackling SSC: 1) direct transformation approaches, which apply direct signal processing operations (such as filtering) over the source speech signal to convert it into target speech, and 2) vocoder-based parametric approaches, which employ a vocoder for feature extraction at the front end of the system and synthesis of the mapped features from source to target speech at the back end of the system. Both approaches have been used in SSC, and each approach has its own advantages. SSC systems following the vocoder-based parametric approach appear generally more flexible than those following the direct-transformation approach. This is because of the parametrization in the system, which allows for selection of the features mapped and therefore adapts the system easily to different conversion tasks and applications. Additionally, interpretability of the results is generally better for vocoder-based parametric SSC methods. On the other hand, while ML-based mapping has been used successfully in direct-transformation SSC methods (such as in [22]), ML-based mapping fits more naturally in a vocoder-based parametric SSC construction. Proof of this can be found in the fact that only one of the direct transformation SSC methods referenced in Section 4 relies on ML-based mapping, while most of the referenced vocoder-based parametric SSC methods use ML-based mapping. ML-based mapping allows for more flexible and adaptive results based on the input data at hand. However, one disadvantage of a vocoder-based parametric SSC system is its heavy reliance on features, which can be erroneously extracted. This can negatively impact the conversion result—for example, the coupling effect of the vocal tract and glottal source is a common issue in many vocoders. Therefore, the choice of vocoder greatly affects the end result, as was discussed earlier. Direct transformation methods may also prove to be less computationally taxing than vocoder-based parametric methods, since the vocoder may utilize computationally expensive algorithms for feature extraction. ML-based mapping can also be computationally expensive, especially

when it is used for several features. Therefore, direct transformation may be the better option in cases where the conversion is fixed (source and target speech in the same style) and where the task requires small time delays.

## 9.4 Discussion of mapping techniques

In the first of our two studies, we also examined an oracle condition, thus bypassing mapping and using the true MFBEs of shouted speech when applying PSM for conversion. The results of the SD condition, when mapping was done using a model trained with speaker-dependent data, were close to those of the oracle condition. Consequently, GMM mapping proved to suffice for the speaker recognition task.

As part of our second study, we evaluated whether the proposed mapping system (BGMM) would function adequately for the given task. The results showed that the BGMMs employed managed to successfully map the selected features from normal to Lombard speech. We hypothesized that the use of BGMMs here would be a step up from the mapping sytem used earlier (GMMs), since BGMMs can cope better with data scarcity. Unfortunately, we cannot compare the mapping models from these two cases since the proposed SSC system and rating tests were different in each case. Nevertheless, the published work presented in that portion of this thesis [27] was extended in [76]. There, GMM and BGMM mappings were compared alongside a feed-forward DNN. Those results showed that DNN mapping yielded a larger degree of conversion, while the GMM and BGMM mappings yielded similar results. However, as noted in [76], BGMMs still provide an advantage over GMMs since the former do not require tuning in order to select the number of clusters.

The two experimental works presented in this thesis propose SSC methods that rely on parallel data. As previously mentioned, due to the scarcity of data on different speaking styles, the work in the second experiment [27] involved feature mapping using a statistical model (BGMM) that is robust and able to handle small data sets [27]. To the best of our knowledge, [27] was the first to employ BGMMs in SSC. BGMMs were utilized earlier for mapping in VC-related research [144]. The work in [27] was extended in [76] by co-authors of the experimental works presented in this thesis. In [76], BGMMs were one of the ML-based mappings under comparison. The encouraging results from the use of BGMM mapping in [27, 76] reinforce the benefits that using a robust mapping technique can bring to SSC, given that data scarcity is a common issue. Thus, a natural step toward advancing high-vocal-effort-focused SSC research would be to use non-parallel mapping techniques to carry out conversion tasks. This step forward has been taken in [91, 95] and by co-authors of the experimental works presented in this thesis.

# References


[1] Z. Inanoglu and S. Young, "A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality," in *Proceedings of Interspeech*, 2007, pp. 490–493.

[2] ——, "Data-driven emotion conversion in spoken English," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[3] D. Erro, E. Navas, I. Hernáez, and I. Saratxaga, "Emotion conversion based on prosodic unit selection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 974–983, 2009.

[4] D. Govind, S. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information," in *Proceedings of Interspeech*, 2011, pp. 2969–2972.

[5] A. K. Vuppala and S. R. Kadiri, "Neutral to anger speech conversion using non-uniform duration modification," in *Proceedings of IEEE International Conference on Industrial and Information Systems*, 2014, pp. 1–4.

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

[7] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 1, pp. 96–103, 1985.

[8] ——, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *Journal of Speech, Language, and Hearing Research*, vol. 29, no. 4, pp. 434–446, 1986.

[9] B. Langner and A. W. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2005, pp. I–265.

[10] À. Calzada Defez and J. C. D. S. Carrié, "Voice quality modification using a harmonics plus noise model," *Cognitive Computation*, vol. 5, no. 4, pp. 473–482, 2013.

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

[12] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.

[13] A. Schmidt-Nielsen and T. H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 249–266, 2000.

[14] A. Alexander, F. Botti, D. Dessimoz, and A. Drygajlo, "The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications," *Forensic Science International*, vol. 146, pp. S95–S99, 2004.

[15] S. S. Kajarekar, H. Bratt, E. Shriberg, and R. De Leon, "A study of intentional voice modifications for evading automatic speaker recognition," in *Proceedings of IEEE Odyssey - The Speaker and Language Recognition Workshop*, 2006, pp. 1–6.

[16] V. Hautamäki, T. Kinnunen, M. Nosratighods, K.-A. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," *Proceedings of*, pp. 1473–1476, 2010.

[17] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, 2015.

[18] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[19] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Proceedings of Interspeech*, 2008.

[20] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *Proceedings of Interspeech*, 2007, pp. 2396–2399.

[21] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertaş, "Speaker identification from shouted speech: Analysis + and compensation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8027–8031.

[22] A. Ramírez López, R. Saeidi, L. Juvela, and P. Alku, "Normal-to-shouted speech spectral mapping for speaker recognition under vocal effort mismatch," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 4940–4944.

[23] E. Jokinen, R. Saeidi, T. Kinnunen, and P. Alku, "Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task," *Computer Speech and Language*, vol. 53, pp. 1–11, 2019.

[24] Y. Stylianou, "Voice transformation: a survey," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 3585–3588.

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

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

[27] A. Ramírez 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 *Proceedings of Interspeech*, 2017, pp. 1363–1367.

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

[29] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 1960.

[30] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[31] T. Bäckström *et al.*, "Linear predictive modelling of speech: constraints and line spectrum pair decomposition," Ph.D. dissertation, Helsinki University of Technology, 2004.

[32] J. L. Flanagan, *Speech analysis: Synthesis and perception*. Springer-Verlag, 1972.

[33] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[34] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 7, pp. 663–671, 1994.

[35] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 1902–1915, 2008.

[36] J. D. Markel and A. Gray, *Linear prediction of speech*. Springer-Velag, 1976.

[37] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

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

[39] S. King, "A beginners' guide to statistical parametric speech synthesis," *The Centre for Speech Technology Research, University of Edinburgh*, 2010.

[40] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[41] 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 *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[42] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[43] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, 2010.

[44] 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 *Proceedings of Interspeech*, 2016, pp. 2473–2477.

[45] G. Degottex, P. Lanchantin, M. Gales, G. Degottex, P. Lanchantin, and M. Gales, "A log domain pulse model for parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 1, pp. 57–70, 2018.

[46] J. P. Cabral, K. Richmond, J. Yamagishi, and S. Renals, "Glottal spectral separation for speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 195–208, 2014.

[47] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[48] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.

[49] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 38, 2014.

[50] 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 *Proceedings of Interspeech*, 2014, pp. 780–784.

[51] Q. Hu, Y. Stylianou, K. Richmond, R. Maia, J. Yamagishi, and J. Latorre, "A fixed dimension and perceptually based dynamic sinusoidal model of speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 6270–6274.

[52] J. D. Markel and A. H. Gray, *Linear Prediction of Speech.* Springer-Verlag, 1980.

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

[54] C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 1, pp. 69–81, 1993.

[55] 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 *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5120–5124.

[56] H. Lane, B. Tranel, and C. Sisson, "Regulation of voice communication by sensory dynamics," *Journal of the Acoustical Society of America*, vol. 47, no. 2B, pp. 618–624, 1970.

[57] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Mal. de L'Oreille et du Larynx*, pp. 101–119, 1911.

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

[59] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Lœvenbruck, "An acoustic and articulatory study of Lombard speech: Global effects on the utterance," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2006, pp. 17–22.

[60] 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 Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 366–378, 2009.

[61] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[62] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[63] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[64] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[65] J. Elliott *et al.*, "Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics," in *Proceedings of Australian Internation Conference on Speech Science and Technology*, 2000, pp. 154–159.

[66] J. M. Pickett, "Effects of vocal force on the intelligibility of speech sounds," *Journal of the Acoustical Society of America*, vol. 28, no. 5, pp. 902–905, 1956.

[67] D. Rostolland, "Intelligibility of shouted voice," *Acta Acustica united with Acustica*, vol. 57, no. 3, pp. 103–121, 1985.

[68] ——, "Acoustic features of shouted voice," *Acta Acustica united with Acustica*, vol. 50, no. 2, pp. 118–125, 1982.

[69] J.-S. Liénard and M.-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *Journal of the Acoustical Society of America*, vol. 106, no. 1, pp. 411–422, 1999.

[70] D. Rostolland, "Phonetic structure of shouted voice," *Acta Acustica united with Acustica*, vol. 51, no. 2, pp. 80–89, 1982.

[71] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, and P. Alku, "Analysis and synthesis of shouted speech," in *Proceedings of Interspeech*, 2013, pp. 1544–1548.

[72] D.-Y. Huang, S. Rahardja, and E. P. Ong, "Lombard effect mimicking," in *Proceedings of ISCA Workshop on Speech Synthesis*, 2010, pp. 258–263.

[73] K. Nathwani, G. Richard, B. David, P. Prablanc, and V. Roussarie, "Speech intelligibility improvement in car noise environment by voice transformation," *Speech Communication*, vol. 91, pp. 17–27, 2017.

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

[75] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998, pp. 285–288.

[76] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, vol. 7, pp. 17 230–17 246, 2019.

[77] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[78] P. Senin, "Dynamic time warping algorithm review," University of Hawaii at Manoa, Technical report, 2008.

[79] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[80] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[81] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1301–1312, 2006.

[82] L. M. Arslan and D. Talkin, "Speaker transformation using sentence HMM based alignments and detailed prosody modification," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1998, pp. 289–292.

[83] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," in *Proceedings of Interspeech*, 2008, pp. 1453–1456.

[84] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2007.

[85] D. Reynolds, *Gaussian Mixture Models*. Springer, 2009, pp. 659–663.

[86] D. Steinberg, "An unsupervised approach to modelling visual data," Ph.D. dissertation, University of Sydney, 2013.

[87] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[88] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," University of British Columbia, Technical report, 2007.

[89] C. M. Bishop, *Pattern recognition and machine learning.* Springer, 2006.

[90] G. Parisi, *Statistical field theory.* Addison-Wesley, 1988.

[91] S. Seshadri, L. Juvela, J. Yamagishi, O. Räsänen, and P. Alku, "Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 6835–6839.

[92] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[93] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2009.

[94] 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 *Proceedings of Interspeech*, 2016, pp. 2771–2775.

[95] S. Seshadri, L. Juvela, P. Alku, and O. Räsänen, "Augmented cycleGANs for continuous scale normal-to-Lombard speaking style conversion," in *Proceedings of Interspeech*, 2019, pp. 2838–2842.

[96] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville, "Augmented cycleGAN: Learning many-to-many mappings from unpaired data," in *Proceedings of International Conference on Machine Learning*, vol. 80, 2018, pp. 195–204.

[97] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Hurricane natural speech corpus [sound]," Website resource: https://datashare.is.ed.ac.uk/handle/10283/347, 2015, accessed 06.03.2020.

[98] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering and Physics*, vol. 24, no. 7-8, pp. 515–520, 2002.

[99] F. Ahmadi, I. V. McLoughlin, and H. R. Sharifzadeh, "Analysis-by-synthesis method for whisper-speech reconstruction," in *Proceedings of IEEE Asia Pacific Conference on Circuits and Systems*, 2008, pp. 1280–1283.

[100] 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 *Proceedings of IEEE International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 374–377.

[101] A. P. Passos, "A lightweight processing for conversion of whispering voice into normal speech," in *Proceedings of IEEE International Conference on Audio, Language and Image Processing*, 2010, pp. 74–79.

[102] C. Huang, X. Y. Tao, L. Tao, J. Zhou, and H. B. Wang, "Reconstruction of whisper in Chinese by modified MELP," in *Proceedings of IEEE International Conference on Computer Science and Education*, 2012, pp. 349–353.

[103] I. V. McLoughlin, J. Li, and Y. Song, "Reconstruction of continuous voiced speech from whispers," in *Proceedings of Interspeech*, 2013, pp. 1022–1026.

[104] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad, "Fundamental frequency generation for whisper-to-audible speech conversion," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 2579–2583.

[105] J.-J. Li, I. V. McLoughlin, L.-R. Dai, and Z.-h. Ling, "Whisper-to-speech conversion using restricted Boltzmann machine arrays," *Electronics Letters*, vol. 50, no. 24, pp. 1781–1782, 2014.

[106] H. Konno, M. Kudo, H. Imai, and M. Sugimoto, "Whisper to normal speech conversion using pitch estimated from spectrum," *Speech Communication*, vol. 83, pp. 10–20, 2016.

[107] G. N. Meenakshi and P. K. Ghosh, "Whispered speech to neutral speech conversion using bidirectional LSTMs," in *Proceedings of Interspeech*, 2018, pp. 491–495.

[108] J. H. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, vol. 5, 1997, pp. 2387–2390.

[109] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[110] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, 1998.

[111] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

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

[113] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, 1995, pp. 451–454.

[114] P. Depalle and T. Helie, "Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[115] X. Fan and J. H. Hansen, "Speaker identification with whispered speech based on modified LFCC parameters and feature mapping," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 4553–4556.

[116] C. Hanilçi, T. Kinnunen, P. Rajan, J. Pohjalainen, P. Alku, and F. Ertas, "Comparison of spectrum estimators in speaker verification: mismatch conditions induced by vocal effort," in *Proceedings of Interspeech*, 2013, pp. 2881–2885.

[117] J. Pohjalainen, C. Hanilçi, T. Kinnunen, and P. Alku, "Mixture linear prediction in speaker verification under vocal effort mismatch," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1516–1520, 2014.

[118] R. Saeidi, P. Alku, and T. Bäckström, "Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 42–53, 2016.

[119] M. Sarria-Paja, M. Senoussaoui, D. O'Shaughnessy, and T. H. Falk, "Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5480–5484.

[120] J. Galić, S. Jovičić, V. Delić, B. Marković, D. Š. Pavlović, and Đ. Grozdić, "HMM-based whisper recognition using $\mu$-law frequency warping," *SPIIRAS Proceedings Journal*, vol. 3, no. 58, pp. 27–52, 2018.

[121] L. Juvela, "Perceptual spectral matching utilizing mel-scale filterbanks for statistical parametric speech synthesis with glottal excitation vocoder," Master's thesis, Aalto University, 2015.

[122] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, pp. 142–152, 2008.

[123] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. Prentice-Hall, 1974.

[124] N. Levinson, "The Wiener (root mean square) error criterion in filter design and prediction," *Journal of Mathematics and Physics*, vol. 25, no. 1-4, pp. 261–278, 1946.

[125] J. Durbin, "The fitting of time-series models," *Revue de l'Institut International de Statistique*, vol. 28, pp. 233–244, 1960.

[126] J. Pohjalainen, T. Raitio, S. Yrttiaho, and P. Alku, "Detection of shouted speech in noise: human and machine," *Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2377–2389, 2013.

[127] R. Saeidi and D. A. V. Leeuwen, "The Radboud University Nijmegen submission to NIST SRE-2012," in *Proceedings of NIST Speaker Recognition Evaluation Workshop*, 2012.

[128] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[129] D. Hardt and K. Fellbaum, "Spectral subtraction and RASTA-filtering in text-dependent HMM-based speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1997, pp. 867–870.

[130] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proceedings of Interspeech*, 2006, pp. 1471–1474.

[131] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[132] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - A unified approach to speech spectral estimation," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 1994, pp. 1043–1046.

[133] J. Kämäräinen and P. Paalanen, "GMMBayes Toolbox V1.0." Website resource: https://www.it.lut.fi/project/gmmbayes/, 2005, accessed 06.03.2020.

[134] D. Ellis, "Dynamic time warp (DTW) in Matlab," Website resource: http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/, 2003, accessed 06.03.2020.

[135] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011–1031, 2000.

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

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

[138] D. Talkin and Google, "REAPER: Robust Epoch And Pitch EstimatoR," Website resource: https://github.com/google/REAPER, 2015, accessed 06.03.2020.

[139] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT Press, 2012.

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

[141] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978.

[142] J. M. Chambers, *Graphical methods for data analysis*. Chapman and Hall/CRC, 1983.

[143] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics*, vol. 31, pp. 1–186, 1980.

[144] L. Li, Y. Nankaku, and K. Tokuda, "A Bayesian approach to voice conversion based on GMMs using multiple model structures," in *Proceedings of Interspeech*, 2011.