Speech Communication 88 (2017) 65–82

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

An overview of voice conversion systems

Seyed Hamidreza Mohammadi*, Alexander Kain

Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR, USA

ARTICLE INFO

Article history:
Received 22 November 2015
Revised 10 January 2017
Accepted 15 January 2017
Available online 16 January 2017

Keywords:
Voice conversion
Overview
Survey

ABSTRACT

Voice transformation (VT) aims to change one or more aspects of a speech signal while preserving linguistic information. A subset of VT, Voice conversion (VC) specifically aims to change a *source* speaker's speech in such a way that the generated output is perceived as a sentence uttered by a *target* speaker. Despite many years of research, VC systems still exhibit deficiencies in accurately mimicking a target speaker spectrally and prosodically, and simultaneously maintaining high speech quality. In this work we provide an overview of real-world applications, extensively study existing systems proposed in the literature, and discuss remaining challenges.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Voice transformation refers to the various modifications one may apply to human-produced speech (Stylianou, 2009); specifically, VT aims to modify one or more aspects of the speech signal while retaining its linguistic information. *Voice conversion* is a special type of VT whose goal is to modify a speech signal uttered by a *source* speaker to sound as if it was uttered by a *target* speaker, while keeping the linguistic contents unchanged (Childers et al., 1989). In other words, VC modifies speaker-dependent characteristics of the speech signal, such as spectral and prosodic aspects, in order to modify the perceived speaker identity while keeping the speaker-independent information (linguistic contents) the same. There is also another class of voice transformations called *voice morphing* where the voices of two speakers are blended to form a virtual third speaker (Cano et al., 2000). VT approaches can be applied to solve related problems, such as changing one emotion into another (Kawanami et al., 2003), improving the intelligibility of speech (Kain et al., 2007), or changing whisper/murmur into speech without modifying speaker identity and linguistic content. For more information regarding applications, please see Section 8. In this work, we will focus on studies pertaining to VC systems, since the majority of important milestones of the VT field have been studied in the VC literature.

An overview of a typical VC system is presented in Fig. 1 (Erro et al., 2010a). In the training phase, the VC system is presented with a set of utterances recorded from the source and target speakers (the training utterances). The speech analysis and mapping feature computation steps encode the speech waveform signal into a representation that allows modification of speech properties. Source and target speakers' speech segments are aligned (with respect to time) such that segments with similar phonetic content are associated with each other. The *mapping* or *conversion* function is trained on these aligned mapping features. In the conversion phase, after computing the mapping features from a new source speaker utterance, the features are converted using the trained conversion function. The speech features are computed from the converted features which are then used to synthesize the converted utterance waveform.

There are various ways to categorize VC methods. One factor is whether they require *parallel* or *non-parallel* recordings during their training phase. Parallel recordings are defined as utterances that have the same linguistic content, and only vary in the aspect that needs to be mapped (speaker identity, in the VC case) (Mouchtaris et al., 2004a). A second factor is whether they are *text-dependent* or *text-independent* (Sündermann et al., 2004b). Text-dependent approaches require word or phonetic transcriptions along with the recordings. These approaches may require parallel sentences recorded from both source and target speakers. For text-independent approaches, there is no transcription available, therefore these approaches require finding speech segments with similar content before building a conversion function (Sündermann, 2008). A third factor is based on the language that source and target speakers speak. Language-independent or *cross-language* VC assumes that source and target speakers speak in different languages (Sündermann et al., 2003; Türk, 2007). Because of the differences in languages, some phonetic classes may not correspond to each other, resulting in problems during mapping. To solve this issue, a combination of non-parallel, text-independent approaches can been used. Another important factor for VC cat-

* Corresponding author.
E-mail addresses: mohammah@ohsu.edu (S.H. Mohammadi), kaina@ohsu.edu (A. Kain).

http://dx.doi.org/10.1016/j.specom.2017.01.008
0167-6393/© 2017 Elsevier B.V. All rights reserved.

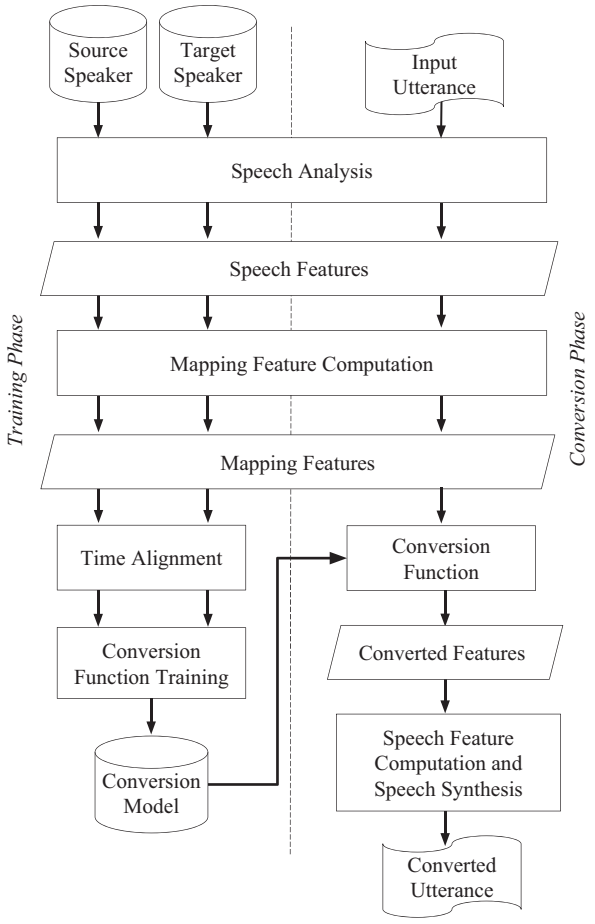

Fig. 1. Training and conversion phases of a typical VC system.

egorization is the amount of the training data that is available. Typically, for larger training data, conversion functions that memorize better are more effective; however, for smaller training data, techniques that generalize better are more preferable.

Some investigators have studied the contributions of several speech features such as of pitch, formant frequencies, spectral envelope and others to speaker individuality (Matsumoto et al., 1973; Kuwabara and Sagisak, 1995). The three most relevant factors were found to be average spectrum, formants, and the average pitch level. As a result, the majority of VC systems aim to modify short-time spectral envelopes and the pitch value. In this study, we present the spectral and prosodic mappings that have been proposed for VC in Sections 5 and 6, respectively. We also review prominent approaches for evaluating the performance of VC systems in Section 7. We then review the different applications that use VC and VT methods in Section 8. Finally, we conclude with reviewing the remaining VC and VT challenges and future directions.

2. Speech features

As shown in Fig. 1, in order to perform voice conversion, analysis/synthesis of the speech signal is necessary. The goal is to extract speech features that allow a good degree of modification with respect to the acoustic properties of speech. Most techniques work on the frame-level (or frame-by-frame), defined as short time segments (~20 ms), in which the length of the frame is chosen so that it satisfies the assumption that the speech signal is stationary (the statistical parameters of the signal over time are fixed) in that frame. The frame can be fixed length throughout the analysis or it can be have a length relative to the pitch periods of the signal (pitch-synchronous analysis).

Speech models can be broadly categorized into source-filter models and signal-based models. In source-filter models, speech is modeled as a combination of a excitation or source signal (representing the vocal cords, not to be confused with the source speaker), and a spectral envelope filter (representing the vocal tract). The model assumes that speech is produced by passing an excitation signal (related to vocal cord movements and frication noise) through the vocal tract (represented by a filter), or, in other words, filtering the excitation signal with the vocal tract filter. The excitation signal and filter are assumed to be independent of each other. Two prominent filter models are commonly used: all-pole and log-spectrum filters. Linear predictive coding (LPC) is an implementation of all-pole models, and mel-log spectrum approximation (MLSA) is an implementation of log-spectrum filters (Imai et al., 1983). SPTK is a publicly available toolkit that provides linear predictive and MLSA analysis/synthesis (Imai et al., 2009). When estimating the spectral envelope, the pitch periods present in the speech signal can show up as harmonics (sharp peaks and valleys) in the spectral envelope. This phenomenon can be problematic when performing any further modifications to the spectrum, since the presence of pitch information in the spectrum would fail the assumption of the independence of source signal and filter. In an attempt to alleviate the interference between signal periodicity and the spectrum, *STRAIGHT* proposes a pitch-adaptive time-frequency spectral smoothing (Kawahara et al., 1999), which was later extended to *TANDEM-STRAIGHT* to provide a unified computation of spectrum, fundamental frequency, and aperiodicity (Kawahara et al., 2008). The advantage of a smooth spectrum is that it provides a representation that is easier to model and manipulate. *CheapTrick* and WORLD propose some improvements over TANDEM-STRAIGHT (Morise, 2015; Morise et al., 2016). The excitation signal can be modeled in various ways. A simple implementation is the pulse/noise model in which the voiced speech segments are modeled using a periodic pulse and the unvoiced speech segments are modeled using noise. More complex excitation signal models such as glottal excitation models (Childers, 1995; Vincent et al., 2007; Del Pozo and Young, 2008; Pozo, 2008; Agiomyrgiannakis and Rosec, 2009), residual signals (Kain and Macon, 2001; Sündermann et al., 2005; Zhang et al., 2005; Duxans and Bonafonte, 2006; Percybrooks and Moore, 2008), mixed excitation (Ohtani et al., 2006; Nurminen et al., 2007), and band aperiodicity (Helander et al., 2012; Chen et al., 2016) have been used.

Signal-based analysis/synthesis approaches model the speech signal by not making any restrictive assumptions (such as the independence of source signal and filter); hence they usually have higher quality. The downside is that they are less flexible for modification. A simple analysis/synthesis technique is pitch-synchronous overlap-add (PSOLA) (Moulines and Charpentier, 1990). PSOLA uses varying frame sizes related to the fundamental frequency (F_0) to create short frames of the signal, keeping the signal in time-domain. PSOLA allows for prosodic transformations of pitch and duration. Linear Predictive PSOLA adds the ability to perform simple vocal tract modifications (Valbret et al., 1992a). Harmonic plus noise models (HNM) assume that the speech signal can be decomposed into harmonics (sinusoids with frequencies relevant to pitch). HNMs generate high quality speech but they are not as flexible as source-filter models for modification, mainly because of the difficulty of dealing with phase (Stylianou, 1996). AHOCODER is a publicly available toolkit that provides high-quality HNM synthesis (Erro et al., 2011). Speech signals can also be represented as a sum of non-stationary modulated sinusoids; this has shown to significantly improve the synthesized speech quality in low-resource settings (Agiomyrgiannakis, 2015).

3. Mapping features

One might directly use speech analysis output features for training the mapping function. More commonly, the speech features are further processed to allow better representation of speech. As shown in Fig. 1, following the speech analysis step, the mapping features are computed from the speech features. The aim is to obtain representations that allow for more effective manipulation of the acoustic properties of speech.

3.1. Local features

Local features represent speech in short-time segments. The following features are commonly utilized to represent local spectral features:

- **Spectral envelope:** the logarithm of the magnitude spectrum can be used directly for representing the spectrum. Because of the high dimensionality of these parameters, more constrained VC mapping functions are commonly used (Valbret et al., 1992a; Sündermann et al., 2003; Mohammadi and Kain, 2013). The frequency scale can be warped to Mel- or Bark-scale, which are frequency scales that emphasize perceptually relevant information. Recently, due to the prevalence of neural network techniques and their ability to handle high-dimensional data, these features are becoming more popular. Spectral parameters have high inter-correlation.
- **Cepstrum:** a spectral envelope can be represented in the cepstral domain using a finite number of coefficients computed by the Discrete Cosine Transform of the log-spectrum. Commonly, mel-cepstrum is used in the literature (Imai, 1983). Mel-cepstrum (MCEP) is a more commonly used variation. Cepstral parameters have low inter-correlation.
- **Line spectral frequencies (LSF):** manipulating LPC coefficients may cause unstable filters, which is the reason that usually LSF coefficients are used for modification. LSFs are more related to frequency (and formant structure), and they also have better quantization and interpolation properties (Paliwal, 1995). These properties make them more appropriate when statistical methods are used (Kain, 2001). LSF parameters have high inter-correlation. These parameters are also known as Line spectral pairs (LSP).
- **Formants:** formant frequencies and bandwidths can be used to represent a simplified version of the spectrum (Mizuno and Abe, 1995; Zolfaghari and Robinson, 1997; Rentzos et al., 2003; Godoy et al., 2010b). They represent spectral features which are of high importance to speaker identity; however, because of their compact nature, they can result in low speech quality during more complex acoustic events.

The local pitch features are typically represented by F_0, or alternatively by logarithm of F_0 which is considered to be more perceptually relevant.

3.2. Contextual features

Most of the mapping functions assume frame-by-frame processing. Human speech is highly dynamic over longer segments and the frame-by-frame assumption restricts the modeling power of the mapping function. Ideally, speech segments with similar static features but different dynamic features should not be treated the same. Techniques that add contextual information to the features are proposed: appending multiple frames, appending delta (and delta-delta) features, and event-based encodings. Appending multiple frames forms a new super-vector feature (Wu et al., 2013d; Chen et al., 2014a; Mohammadi and Kain, 2015) on which the mapping function is trained. This new multi-frame feature would allow the mapping function to capture the transitions within the short (but longer than a single frame) segments, since the number of neighboring frames that are appended is chosen in a way that meaningful transitional information is present within the segment. In another approach, appending delta and delta-delta features has been proposed (Furui, 1986); this allows the mapping function to also consider the dynamic information in the training phase (Duxans et al., 2004). Moreover, during computing speech features from the converted features, this dynamic information can be utilized to generate a local feature trajectory that considers both static and dynamic information (Toda et al., 2007a). Event-based approaches decompose local feature sequence into event targets and event transitions to effectively model the speech transition. Temporal decomposition (TD) decomposes local feature sequence into event targets and event functions (Nguyen and Akagi, 2007; 2008; Nguyen, 2009). The event functions connect the event targets through time. Similarly, Asynchronous interpolation model (AIM) proposes to encode local feature sequence by a set of basis vectors and connection weights (Kain and van Santen, 2007). The connection weights connect the basis vectors through time to model feature transition. The main difficulty with the event-based approaches is to correctly identify event locations in the sequence.

Analogous to spectral parameterization, contextual information can be added to the local pitch features as well. More meaningful speech units such as syllables can be considered to encode contextual information. We present pitch parametrization and mapping approaches in more detail in Section 6.

In addition to these techniques that explicitly encode the speech dynamics, some mapping functions implicitly model dynamics from a local feature sequence. Examples of these implicit dynamic models are hidden Markov models (HMMs) and recurrent neural networks (RNNs). These models typically encompass a concept of *state*. The state that the model is currently in is determined by the previously seen samples in the sequence, hence allowing the model to capture context. We will mention these approaches at the end of their relevant spectral mapping subsections in Section 5.

4. Time-alignment

As shown in Fig. 1, VC techniques commonly utilize parallel source-target feature vectors for training the mapping function between source and target features. The most common approach uses recordings of a set of *parallel* sentences (sentences including the same linguistic contents) from both source and target speakers. However, the source and target speakers are likely to have different-length recordings, and have dissimilar phoneme durations within the utterance as well. Therefore, a time-alignment approach must be used to address the temporal differences. Manual or automatic phoneme transcriptions can be utilized for time alignment. Most often, a dynamic time warping (DTW) algorithm is used to compute the best time alignment between each utterance pair (Abe et al., 1988; Kain and Macon, 1998a), or within each phoneme pair. The final product of this step is a pair of source and target feature sequences of equal length. The DTW alignment strategy assumes that the same phonemes of the speakers have similar features (when using a particular distance measure). This assumption however is not always true and might result in sub-optimal alignments, since the speech features are typically not speaker-independent. For improving the alignment output, one can iteratively perform the alignment between the target features and the converted features (instead of source features), followed by training and conversion, until a convergence condition is satisfied. There are various methods that perform time alignment in different conditions, depending on the availability of parallel

Table 1
Overview of time-alignment methods for VC.

Method	Parallel recording	Phonetic transcription	Cross-language	Implicit in training
DTW (Abe et al., 1988)	yes	no	no	no
DTW including phonetics (Kain and Macon, 1998a)	yes	yes	no	no
Forced alignment (Arslan and Talkin, 1998; Ye and Young, 2006)	yes	Forced alignment	no	no
Time sequence matching (Nankaku et al., 2007)	yes	no	no	yes
TTS with same duration (Duxans et al., 2006; Wu et al., 2006)	no	yes	no	no
ASR-TTS with same duration (Ye and Young, 2004; Tao et al., 2010)	no	ASR	no	no
Model alignment (Zhang et al., 2008)	no	no	yes	yes
Unit-selection alignment (Arslan and Talkin, 1998; Sündermann and Ney, 2003; Erro and Moreno, 2007a; Sündermann et al., 2004a)	no	no	yes	no
Iterative (INCA) (Erro and Moreno, 2007a; Erro et al., 2010a)	no	no	yes	no
Unit-selection VC (Sündermann et al., 2006a, c)	no	no	yes	yes
Model adaptation (Mouchtaris et al., 2006; Lee and Wu, 2006)	no	no	no	yes

recordings, the availability of phonetic transcription, the language of the recordings, and whether the alignment is implicit in training or is performed separately. An overview of some time-alignment methods is given in Table 1.

More complicated approaches are required for non-parallel alignment. One set of alignment methods use transcribed, non-parallel recordings for training purposes. For alignment, a unit-selection text-to-speech (TTS) system can be used to synthesize the same sentences for both source and target speakers (Duxans et al., 2006). The resulting speech is completely aligned, since the duration of the phonemes can be specified to the TTS system beforehand (Wu et al., 2006). These approaches usually require a relatively large number of training utterances and they are usually more suited for adapting an already trained parametric TTS system to new speakers/styles. These approaches, however, are text-dependent. For text-independent, non-parallel alignment, a unit-selection approach that selects units based on input source features is proposed to select the best-matching source-target feature pairs (Sündermann et al., 2006a). The INCA algorithm (Erro and Moreno, 2007a; Erro et al., 2010a) iteratively finds the best feature pairs between the converted source and the target utterances using a nearest neighbors algorithm, and then trains the conversion on those pairs. This process is iterated until the converted source converges and stops changing significantly.

Researchers have studied the impact of frame alignment on VC performance, specifically the situation where one frame aligns with multiple other frames (hence making the source-target feature relationship not one-to-one), and approaches to reduce the resulting effects were proposed (Mouchtaris et al., 2007; Helander et al., 2008b; Godoy et al., 2009); notably, some studies suggested to filter out the source-target training pairs that are unreliable, based on a confidence measure (Turk and Arslan, 2006; Rao et al., 2016).

5. Spectral modeling

This section discusses the mappings that are used for VC task to learn the associations between the spectral mapping features. We assume that the mapping features are aligned using one of the techniques described in Section 4. In addition, we assume that the training source and target speaker features are sequences of length N represented by $\mathbf{X}^{train} = [\mathbf{x}_1^{train}, \ldots, \mathbf{x}_N^{train}]$ and $\mathbf{Y}^{train} = [\mathbf{y}_1^{train}, \ldots, \mathbf{y}_N^{train}]$, respectively, where each element is a D-dimensional vector $\mathbf{x}^\top = (x_1, \ldots, x_D)$. Each element of the sequence represents the feature computed in a certain frame, where the features can be any of the mapping features described in Section 3. The goal is to build a feature mapping function $\mathcal{F}(\mathbf{X})$ that maps the source feature sequence to be more similar the target speaker feature sequence, as shown in Eq. (1). At conversion time, an unseen source feature $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N^{test}}]$ of length N^{test} will be passed to the function in order to predict target features,

$$\mathcal{F}(\mathbf{X}) = \hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_{N^{test}}] \tag{1}$$

Traditionally we assume that the mappings are performed frame-by-frame, meaning that each frame is mapped independent of other frames,

$$\hat{\mathbf{y}} = \mathcal{F}(\mathbf{x}) \tag{2}$$

however, more recent models consider more context to go beyond frame-by-frame mapping, which are mentioned at the end of their relevant subsections.

In Fig. 2, we devise a toy example to show the performance of some conversion techniques. We utilize 40 sentences from a male (source) and a female (target) speaker from the Voice Conversion Challenge corpus (refer to Section 7). We extract 24th-order MCEP features and use principal component analysis (PCA) on both speaker's data to reduce the dimensionality to two for easier two-dimensional visualization. The yellow and green dots represent source and target training features. The input data, represented as magenta, is a grid over the source data distribution in the top row, and the feature sequence of a word uttered by the source speaker (excluded from the training data) in the bottom row. The original target and converted features are represented as blue and red, respectively.

5.1. Codebook mapping

Vector quantization (VQ) can be used to reduce the number of source-target pairs in an optimized way (Abe et al., 1988). This approach creates M code vectors based on hard clustering using vector quantization on source and target features separately. These code vectors are represented as $\mathbf{c}_m^x$ and $\mathbf{c}_m^y$ for source and target speakers, for $m = [1, \ldots, M]$, respectively. At conversion time, the closest centroid vector of the source codebook is found and the corresponding target codebook is selected

$$\mathcal{F}_{\text{VQ}}(\mathbf{x}) = \mathbf{c}_m^y, \tag{3}$$

where $m = \arg_{\eta=[1,M]} \min d(\mathbf{c}_\eta^x, \mathbf{x})$. The VQ approach is compact and covers the acoustic space appropriately since a clustering approach is used to determine the codebook. However, this simple approach still has the disadvantage of generating discontinuous feature sequences. This phenomenon can be solved by using a large M but this requires a large amount of parallel-sentence utterances. The quantization error can be reduced by using a fuzzy VQ, which uses soft clustering (Shikano et al., 1991; Arslan and Talkin, 1997; Turk and Arslan, 2006). For an incoming new source mapping feature, a continuous weight w_m^x is computed for each codebook based on a weight function. The mapped feature is calculated as a weighted sum of the centroid vectors

$$\mathcal{F}_{\text{fuzzy VQ}}(\mathbf{x}) = \sum_{m=1}^{M} w_m^x \mathbf{c}_m^y, \tag{4}$$

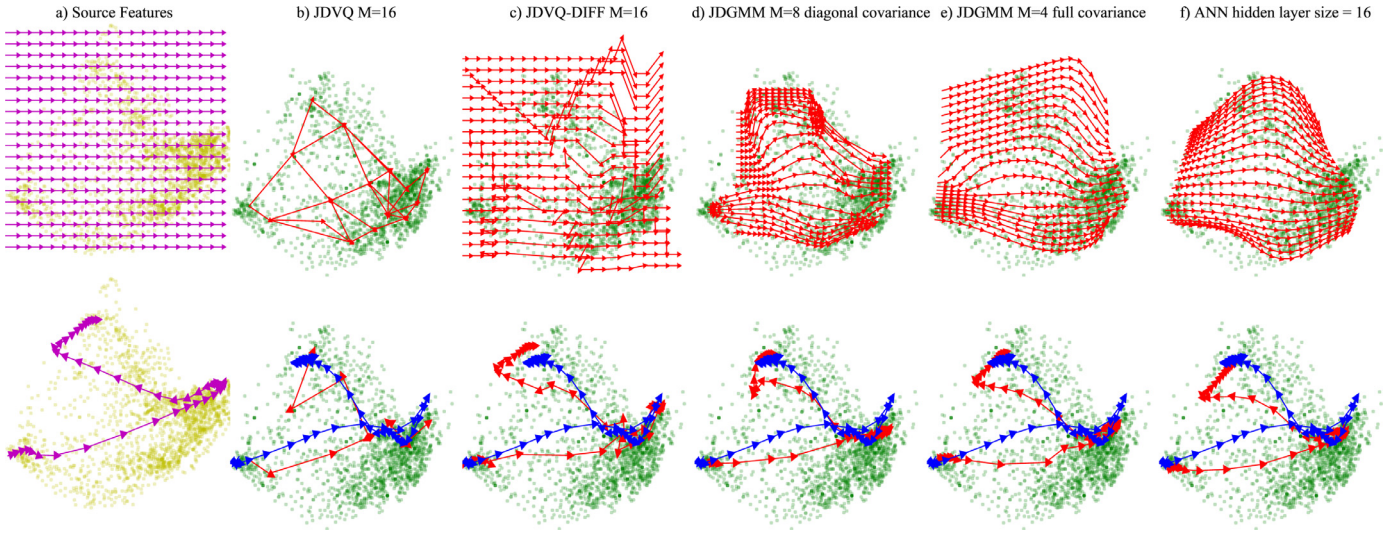

Fig. 2. A toy example comparing JDVQ, JDVQ-DIFF, JDGMM, and ANN. The x- and y-axis are first and second dimensions of PCA, respectively. Color codes for source, target, input, original target, and converted samples are represented as yellow, green, magenta, blue, and red, respectively. The top row shows an example with a grid as input and the bottom row shows an example with a real speech trajectory as input. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where $w_m^x = weight(\mathbf{c}_m^x, \mathbf{x}^{new})$. This weight function can be computed using various methods, including Euclidian distance (Shikano et al., 1991), phonetic information (Shuang et al., 2004), exponential decay (Arslan, 1999), vector field smoothing (Hashimoto and Higuchi, 1995), and statistical approaches (Lee, 2007). Simple VQ is a special case of fuzzy-VQ in which only one of the vectors is assigned the weight value of one, and the rest have zero contribution.

Alternatively, to allow the model to capture more variability and reduce quantization error, a difference vector between the source and target centroids can be stored as codebook (VQ-DIFF) and added to the incoming mapping feature (Matsumoto and Yamashita, 1993)

$$\mathcal{F}_{\text{VQ-DIFF}}(\mathbf{x}) = \mathbf{x} + (\mathbf{c}_m^y - \mathbf{c}_m^x). \tag{5}$$

Similar to fuzzy-VQ, a soft-clustering extension can be applied. For associating the source and target codebooks vectors, the joint-density (JD) can be modeled, in which the source and target vectors are first stacked and then the joint codebook vectors are estimated using the clustering algorithm. As a result, the computed source-target codebook vectors will be associated together. In Fig. 2b and c JDVQ and JDVQ-DIFF conversions are applied to the toy example data. As can be seen in the figure, the JDVQ-DIFF is able to generate samples that were not present in the target training data, however, JDVQ can not make this extrapolation. JDVQ exhibits high quantization error. Both JDVQ and JDVQ-DIFF are prone to generating discontinuous feature sequences.

5.2. Mixture of linear mappings

Valbret et al. (1992a) proposed to use linear multivariate regression (LMR) for each code vector. In this approach, the linear transformation is calculated based on a hard clustering of the source speaker space

$$\mathcal{F}_{\text{LMR}}(\mathbf{x}) = \mathbf{A}_m\mathbf{x} + \mathbf{b}_m, \tag{6}$$

where $m = \arg_{\eta=[1,M]} \min d(\mathbf{c}_\eta^x, \mathbf{x})$, and $\mathbf{A}_m$ and $\mathbf{b}_m$ are regression parameters. This method, however, suffers from discontinuities in the output when the clusters change between neighboring frames. To solve this issue, an idea similar to fuzzy-VQ is proposed, but for linear regression. The previous equation then changes to

$$\mathcal{F}_{\text{weighted LMR}}(\mathbf{x}) = \sum_{m=1}^{M} w_m^x(\mathbf{A}_m\mathbf{x} + \mathbf{b}_m), \tag{7}$$

where $w_m^x = weight(\mathbf{c}_m^x, \mathbf{x})$. Various approaches have been proposed to estimate the parameters of the mapping function. Kain and Macon (1998a) proposed to estimate the joint density of the source-target mapping feature vectors in an approach called joint-density Gaussian mixture model (JDGMM). A joint feature vector $\mathbf{z}_t = [\mathbf{x}_t^\top, \mathbf{y}_t^\top]^\top$ is created, and a Gaussian mixture model (GMM) is fit to the joint data. The parameters of the weighted linear mapping are estimated as

$$\mathbf{A}_m = \mathbf{\Sigma}_m^{xy}\mathbf{\Sigma}_m^{xx-1}, \mathbf{b}_m = \mu_m^y - \mathbf{A}_m\mu_m^x, w_m^x = P(m|\mathbf{x}^{new}), \tag{8}$$

where $\mathbf{\Sigma}_m^{xy}$, $\mathbf{\Sigma}_m^{xx}$, μ_m^x, μ_m^y, and $P(m|\mathbf{x})$ are the mth training cross-covariance matrix, source covariance matrix, source mean vector, target mean vector, and conditional probability of cluster m given input $\mathbf{x}$, respectively. Stylianou et al. (1998) proposed a similar formulation as Eq. (7), however the GMM mixture components are estimated on source feature vectors only, rather than the joint feature vectors. Additionally, instead of computing the cross-covariance matrix and the target means directly from the joint data, they are computed by solving a matrix equations to minimize the least squares via

$$\mathbf{A}_m = \mathit{\Gamma}_m\mathbf{\Sigma}_m^{xx-1}, \mathbf{b}_m = \mathbf{v}_m - \mathbf{A}_m\mu_m^x, w_m^x = P(m|\mathbf{x}^{new}), \tag{9}$$

where $\mathit{\Gamma}$ and $\mathbf{v}$ are the mapping function parameters which are estimated by solving a least squares optimization problem. In the case of JDGMM, $\mathit{\Gamma} = \mathbf{\Sigma}_m^{xy}$ and $\mathbf{v} = \mu_m^y$, which are computed from the joint distribution. JDGMM has the advantage of considering both the source and the target space during training, giving opportunity for more judicious allocation of individual components. Furthermore, the parameters of the conversion function can be directly estimated from the joint GMM and thus a potentially very large matrix inversion problem can be avoided. The derivation of the mapping function parameters are derived similar to Eq. (8). GMM approaches are compared in (Mesbahi et al., 2007a). In Fig. 2d and e, the JDGMM conversion for $M = 8$ with diagonal covariance and $M = 4$ with full covariance matrices are applied to the toy example data, respectively. Both approaches result in smoother

trajectories compared to JDVQ methods. The full covariance matrix seems to capture the distribution of the target speaker better.

One major disadvantage of GMMs is the requirement of computing covariance matrices (Mesbahi et al., 2007a). If we assume a full covariance matrix, the number of parameters is on the order of m multiplied by the square of the dimension of the features. If we do not have sufficient data (which is usually the case in VC), the estimation might result in *over-fitting*. To overcome this issue, diagonal covariance matrices are commonly used in the literature. Due to the assumption of independence between the individual vector components, diagonal matrices might not be appropriate for some mapping features such as LSFs or the raw spectrum. To propose a middle ground between diagonal and full covariance matrices, some studies use a mixture of factor analyzers, which assumes that the covariance structure of the high-dimensional data can be represented using a small number of latent variables (Uto et al., 2006). There also exists an extension of this approach that utilizes non-parallel a priori data (Wu et al., 2012). Another study proposes to use partial least squares (PLS) regression in the transformation (Helander et al., 2010b). PLS is a technique that combines principles from principal component analysis (PCA) and multivariate regression (MLR), and is most useful in cases where the feature dimensionality of $\mathbf{x}_t^{train}$ and $\mathbf{y}_t^{train}$ is high and the features exhibit multicollinearity. The underlying assumption of PLS is that the observed variable $\mathbf{x}_t^{train}$ is generated by a small number of latent variables $\mathbf{r}_t$ which explain most of the variation in the target $\mathbf{y}_t^{train}$, in other words $\mathbf{x}_t^{train} = \mathbf{Q}\mathbf{r}_t + \mathbf{e}_t^x$ and $\mathbf{y}_t^{train} = \mathbf{P}\mathbf{r}_t + \mathbf{e}_t^y$, where $\mathbf{Q}$ and $\mathbf{P}$ are speaker specific transformation matrices and $\mathbf{e}_t^x$ and $\mathbf{e}_t^y$ are residual terms. Solving $\mathbf{Q}$ and $\mathbf{P}$, and extending the model to handle multiple weighted regressions, result in the computation of regression parameters $\mathbf{A}_m$, $\mathbf{b}_m$, and w_m^x, as detailed in (Helander et al., 2010b). The approach is later extended to use kernels and dynamic information, in order to capture non-linear relationships and time-dependencies (Helander et al., 2012).

Various other approaches to estimate regression parameters have been proposed. In the Bag of Gaussian model (BGM) (Qiao et al., 2011), two types of distributions are present. The basic distributions are GMMs, but the approach also uses some complex distributions to handle the samples that are far from the center of their distribution. Other approaches based on Radial Basis Functions (RBFs) (Watanabe et al., 2002; Nirmal et al., 2013) and Support vector regression (SVR) (Laskar et al., 2009; Song et al., 2011) have also been proposed; these use non-linear kernels (such as Gaussian or polynomial) to transform the source mapping features to a high-dimensional space, followed by one linear mapping in that space. Finally, some approaches are physically motivated mappings (Ye and Young, 2003; Zorilă et al., 2012) and local linear transformations (Popa et al., 2012).

One effect of over-fitting, mentioned earlier, is the presence of discontinuity in the generated features. For example, if the number of parameters is high, the converted feature sequence might be discontinuous. For solving this phenomenon, post-filtering of the posterior probabilities (Chen et al., 2003) or the generated features themselves (Toda et al., 2007a; Helander et al., 2010b) has been proposed. Another known effect of GMM-based mappings is generating speech with a muffled quality. This is due to averaging features that are not fully interpolable, which results in wide formant bandwidths in the converted spectra. For example, LSF vectors can use different vector components to track the same formant, and thus averaging across such vectors produces vectors that do not represent realistic speech. This problem is also known as *over-smoothing*, since the converted spectral envelopes are typically smoothened to a degree where important spectral details become lost. The problem can be seen in Fig. 2c where the predicted samples fall well within the probability distribution of the target features and fail to move to the edges of the distribution, thus failing to capture the variability of the target features. To solve this issue, some studies have proposed to post-process the converted features. A selection of post-processing techniques is given in Table 2.

Another framework for solving the VC problem is to view it as a noisy channel model (Saito et al., 2012). In this framework, the output is computed from the conditional maximum-likelihood $\mathcal{F}_{\text{noisy-channel}}(\mathbf{x}) = \texttt{argmax}_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})$, where the conditional probability is defined using Bayes' rule $P(\mathbf{y}|\mathbf{x}) = P(\mathbf{x}|\mathbf{y})P(\mathbf{y})$. The conditional probability $P(\mathbf{x}|\mathbf{y})$ represents the channel properties and is trained on the parallel source-target data, whereas $P(\mathbf{y})$ represents the target properties and is trained on the non-parallel target speaker data. Finally, the problem reduces to decoding of the target features given the observed features, the channel properties, and the target properties. In another framework, the idea of separating style from content is explored using bilinear models (Popa et al., 2009; 2011). For the VC task, style is the speaker identity and content is the linguistic content of the sentence. In this method, two linear mappings are performed, one for style and one for content. During conversion, the speaker identity information of the input utterance is replaced with the target speaker identity information computed during training.

In order to better model dynamics of speech, various approaches such as HMMs have been proposed (Kim et al., 1997; Duxans et al., 2004; Yue et al., 2008; Zhang et al., 2009). These approaches consider some context when decoding the HMM states but the final conversion is usually performed frame-by-frame. Another approach is to append dynamic features (delta and delta-delta, i. e. velocity and acceleration, respectively (Furui, 1986)) to the static features (Duxans et al., 2004), as described in Section 3. A very prominent approach called maximum likelihood parameter generation (MLPG) (Tokuda et al., 1995) has been used for generating feature trajectory using dynamic features (Toda et al., 2007a). MLPG can be used as a post-processing step of a JDGMM mapping. It generates a sequence with maximum likelihood criterion given the static features, the dynamic features, and the variance of the features. This approach is usually coupled with GV to increase the variance of the generated feature sequence. Ideally, MLPG needs to consider the entire trajectory of an utterance to generate the target feature sequence. This property is not desirable for real-time applications. Low-delay parameter generation algorithms without GV (Muramatsu et al., 2008) and with GV (Toda et al., 2012a) have also been proposed. Recently, considering the modulation spectrum of the converted feature trajectory (as a feature correlated with over-smoothing) has been proposed, which resulted in significant quality improvements (Takamichi et al., 2015). Incorporating parameter generation into the training phase itself has also been studied (Zen et al., 2011; Erro et al., 2016).

5.3. Neural network mapping

Another group of VC mapping approaches use artificial neural networks (ANNs). ANNs consist of multiple layers, each performing a (usually non-linear) mapping of the type $\mathbf{y} = f(\mathbf{W}\mathbf{x} + \mathbf{b})$ where $f(\cdot)$ is called the activation function that can be implemented as a sigmoid, tangent hyperbolic, rectified linear units, or linear function. A shallow (two-layered) ANN mapping can be defined as

$$F_{ANN}(\mathbf{x}) = f_2(\mathbf{W}_2 f_1(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2)), \tag{10}$$

where W_i, b_i, and f_i represent the weight, bias and activation function for the ith layer, respectively. ANNs with more than two layers are typically called deep neural networks (DNNs) in the literature. The input and output size are usually fixed depending on the application. (For VC, the input and output size are the source and target mapping feature dimensions.) However, the size of the middle layer and activation function are chosen depending on the experiment and data distributions. The first layer activation function

Table 2
Post-processing techniques for reducing the over-smoothing.

Method	Description
Global variance(GV) (Toda et al., 2005; Benisty and Malah, 2011; Hwang et al., 2013)	Adjusts the variance of generated features to match that of target's
ML parameter generation (Toda et al., 2007a)	Maximizes the likelihood during parameter generation using dynamic features
MMI parameter generation (Hwang et al., 2012)	Maximizes the mutual information during parameter generation using dynamic features
Modulation spectrum (Takamichi et al., 2014)	Adjusts the spectral shape of the generated features
Monte Carlo (Helander et al., 2010a)	Minimizing the conversion error and the sequence smoothness together
L2-norm (Sorin et al., 2011)	Sharpens the formant peaks in spectrum
Error compensation (Villavicencio et al., 2015)	Models error and compensate for it
Residual addition (Kang et al., 2005)	Maps the envelope residual and adds it to the GMM-generated spectrum

is almost always non-linear and the activation function of the last layer is linear or non-linear, depending on the design. If the last layer is linear, the ANN approach can be viewed as an LMR approach, with the difference that the linear regression is applied on a data space that is mapped non-linearly from the mapping feature space, and not directly on the mapping features (similar to RBF and SVR). The weights and biases can be estimated by minimizing an objective function, such as mean squared error, perceptual error (Valentini-Botinhao et al., 2015), or sequence error (Xie et al., 2014a).

ANNs are a very powerful tool, but the training and network design is where most care needs to be exercised since the training can easily get stuck in local minima. In general, both GMMs and ANNs are universal approximators (Titterington et al., 1985; Hornik et al., 1989). The non-linearity in GMMs stems from forming the posterior-probability-weighted sum of class-based linear transformations. The non-linearity in ANNs is due to non-linear activation functions. Laskar et al. (2012) compare ANN and GMM approaches in the VC framework in more detail. In Fig. 2f, the ANN conversion for a hidden layer of size 16 is applied to the toy example data. The ANN trajectory is performing similar to JDGMM with full covariance matrix, which is expected since both are universal approximators.

The very first attempt for using ANNs utilized formant frequencies as mapping features (Narendranath et al., 1995), i. e. the source speaker's formant frequencies were transformed towards target speaker's formant frequencies using a ANN followed by a formant synthesizer. Later, Makki et al. (2007) successfully mapped a compact representation of speech features using ANNs. A more typical approach used a three-layered ANN to map mel-cepstral features directly (Desai et al., 2010). Various other ANN architectures have been used for VC (Ramos, 2016): Feedforward architectures (Desai et al., 2010; Azarov et al., 2013; Liu et al., 2014; Mohammadi and Kain, 2014; Nirmal et al., 2014), restricted Boltzmann machines (RBMs) and their variations (Chen et al., 2013; Wu et al., 2013a; Nakashika et al., 2015a), joint architectures (Chen et al., 2013; Mohammadi and Kain, 2015; 2016), and recurrent architectures (Nakashika et al., 2015b; Sun et al., 2015).

Traditionally, DNN weights are initialized randomly; however, it has been shown in the literature that deep architectures do not converge well due to a vanishing gradient and the likelihood of being stuck in a local minimum solution (Glorot and Bengio, 2010). A regularization technique is typically used to solve this issue. One solution is pre-training the network. DNN training converges faster and to a better-performing solution if their initial parameter values are set via pre-training instead of random initialization (Erhan et al., 2010). This especially important for the VC task since the amount of training data is typically smaller compared to other tasks such as ASR or TTS. Stacked RBMs are used to build speaker-dependent representations of cepstral features for source and target speakers before DNN training (Nakashika et al., 2013; 2014b; 2015c). Similarly, layer-wise generative pre-training using RBMs for VC has been proposed (Chen et al., 2014a; 2014b). Mohammadi and Kain (2014) proposed a speaker-independent pre-training of the DNN using multiple speakers other than source and target speakers using de-noising stacked autoencoders. This approach is later extended to use speakers that sound similar to source and target speakers to pre-train the DNN using joint-autoencoders (Mohammadi and Kain, 2015). In a related study, using multiple speakers as source for training the DNN was proposed (Liu et al., 2015). Alternatively, other techniques such as dropout (Srivastava et al., 2014) and using rectified linear units (Glorot et al., 2011) have shown to be successful.

For capturing more context, Xie et al. (2014a) proposed a sequence error minimization instead of a frame error minimization to train a neural network. The architecture of RNNs allows the network to learn patterns over time. They implicitly model temporal behavior by considering the previous hidden layer state in addition to the current frame (Nakashika et al., 2014a; 2015b; Sun et al., 2015; Ming et al., 2016).

5.4. Dictionary mapping

One of the simplest mapping functions is a look-up table that has source features as entry keys and target features as entry values. For an incoming feature, the function will look up to find the most similar key based on a distance criterion, e. g. an objective distortion measure $d(\cdot)$ similar to one described in Section 7.1. In other words, it will look for the nearest neighbor of the incoming source feature and select its corresponding entry value

$$\mathcal{F}_{\text{lookup}}(\mathbf{x}) = \mathbf{y}_t^{train}, \tag{11}$$

where $t = \arg_{\tau=[1,T]} \min d(\mathbf{x}_\tau^{train}, \mathbf{x})$. A major concern is that the similarity of the source feature does not necessarily guarantee similarity in neighboring target features. This phenomenon will cause discontinuities between the generated target parameter sequence. One approach to overcome the discontinuity of the target feature sequence is to assign a weight to all target feature vectors (computed based on the new source feature vector), which will generate a smoother feature sequence. This category of approaches is called *exemplar-based* VC in the literature (Wu et al., 2014b; Aihara et al., 2014a; Wu et al., 2014a; Aihara et al., 2014b) and the mapping function is given by

$$\mathcal{F}_{\text{exemplar}}(\mathbf{x}) = \sum_{t=1}^{T} w_t^x \mathbf{y}_t^{train}, \tag{12}$$

with $w_t^x = \omega(\mathbf{x}_t^{train}, \mathbf{x})$, where $\omega(\cdot)$ can potentially be any distortion measure. A generic objective distortion measure might result in over-smoothing, since many frames may be assigned non-zero weights and will thus be averaged (unless the mapping features are completely interpolable). Commonly, non-negative matrix factorization (NMF) techniques have been used to compute sparse weights. The goal of NMF is to compute an activation matrix **H**

which represents how well we can reconstruct $\mathbf{x}$ by a non-negative weighted addition of all $\mathbf{x}_t^{train}$ vectors, such that $\mathbf{X} = \mathbf{X}^{train}\mathbf{H}$. The activation matrix $\mathbf{H}$ is calculated iteratively (Wu et al., 2014b). NMF computes a non-negative weight for each entry in the table, which results in the mapping function

$$\mathcal{F}_{\text{NMF}}(\mathbf{X}) = \mathbf{Y}^{train}\mathbf{H}. \tag{13}$$

This relatively sparse weighting over all vectors results in smooth generated feature sequences while reducing over-smoothing. This approach however has the disadvantage of computational complexity, which might not be suitable for some applications. To address this issue, computing the activation matrix in a more compact dimension has been proposed (Wu et al., 2014b). The NMF methods are also inherently well-suited for noisy environments (Takashima et al., 2012; 2013; Masaka et al., 2014). Several other extensions of NMF approaches have been proposed, such as mapping the activation matrix (AIHARA et al., 2015), many-to-many VC (Aihara et al., 2015), including contextual information (Wu et al., 2013c; 2014b; Benisty et al., 2014), and local linear embeddings (Wu et al., 2016).

Another approach to combat discontinuities in the generated features is to take the similarity of the target feature sequence into consideration by using a unit-selection (US) paradigm. US approaches make use of a target cost (similar to a table look-up distortion measure) and a concatenation cost (to ensure the neighboring target features are most similar to each other). Since the units are frames, this method is also referred to as frame-selection (FS). The goal is to find be best sequence of indices of training target vectors $\mathbf{S} = [s_1, \ldots, s_N]$ which minimizes the following cost function (Salor and Demirekler, 2006; Uriz et al., 2008; Lee, 2014):

$$\begin{aligned}\mathcal{F}_{\text{FS}}(\mathbf{X}) = \quad & \arg_{\mathbf{S}=[s_1,\ldots,s_{N^{test}}]} \min \\ & \sum_{n=1}^{N^{test}} \alpha \cdot target(\mathbf{x}_{s_n}, \mathbf{x}_n^{new}) + \\ & (1-\alpha) \cdot concatenation(\mathbf{y}_{s_n}, \mathbf{y}_{s_{n-1}})\end{aligned} \tag{14}$$

where α is used for adjusting the tradeoff between fitting accuracy and the spectral continuity criterion. Since there is an exponential number of permutation of index sequences, a dynamic programming approach such as Viterbi is used to find the optimal target sequence. This can be used for aligning frames before any other type of training, or directly used as a mapping function.

The US/FS approach can be adjusted to create text-independent, non-parallel VC systems (Sündermann et al., 2006a, c). In this variation, a vector $\tilde{\mathbf{x}}_n^{cmp}$ is compared to a target training vector in the dictionary to compute the target cost

$$\begin{aligned}\mathcal{F}_{\text{US}}(\mathbf{X}) = \quad & \arg_{\mathbf{S}=[s_1,\ldots,s_{N^{test}}]} \min = \\ & \sum_{n=1}^{N^{test}} \alpha \cdot target(\mathbf{y}_{s_n}, \tilde{\mathbf{x}}_n^{cmp}) + \\ & (1-\alpha) \cdot concatenation(\mathbf{y}_{s_n}, \mathbf{y}_{s_{n-1}})\end{aligned} \tag{15}$$

where $\tilde{\mathbf{x}}_n^{cmp}$ is either the input source vector (in the absence of any parallel data) (Sündermann et al., 2006a) or a naive conversion to target (in the presence of real or artificial parallel data) (Dutoit et al., 2007). As mentioned in Section 4, these techniques can be used for parallelizing the training data as well.

Combinations and variants of US/FS approaches combined with other mapping approaches have been proposed, such as: dictionary mapping (Fujii et al., 2007), codebook mapping (Kim et al., 1997; Eslami et al., 2011), frequency warping (Shuang et al., 2008; Uriz et al., 2009a), GMM mapping (Duxans et al., 2004), segmental GMM (Gu and Tsai, 2014), k-histogram (Uriz et al., 2009b), exemplar-based VC (Wu et al., 2013c), and grid-based approximation (Benisty et al., 2014). These approaches have some limitations, specifically they can generate discontinuous features. Helander et al. (2007) studied the coverage of speech features when using FS as a mapping for VC, and concluded that a small number of training utterances (which is typical in VC tasks) is not adequate for representing the speaker space.

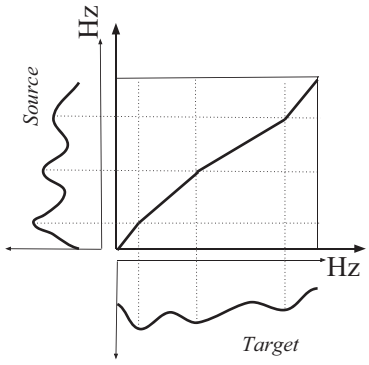

Fig. 3. Piece-wise linear frequency warping function.

5.5. Frequency warping mappings

The estimation of linear regression parameters described in Section 5.2 is typically unconstrained; this can lead to over-fitting. There exist a class of constrained mapping methods which are physically motivated (Zorilă et al., 2012). One common motivation is that two different speakers have different formant frequencies and bandwidths, and different energies in each frequency band. Thus, for conversion, a constrained mapping only allows manipulation of formant location/bandwidths and energy in certain frequency bands. This reduces over-fitting, while allowing the use of high-dimensional mapping features, which is beneficial since vocoders that utilize high-dimensional speech features (e. g. harmonic vocoders) usually have higher speech quality compared to more compact vocoders (e. g. LSF vocoders). The motivation behind the frequency warping (FW) methods is that a mapping of a source speaker spectrum to a target speaker spectrum can be performed by warping the frequency axis, to adjust the location and bandwidth of the formants, and applying amplitude scaling, to adjust the energy in each frequency bands (Erro and Moreno, 2007b; Erro et al., 2010b; Godoy et al.,2012); this approach is more physically interpretable than an unconstrained mapping. Although these approaches can be implemented as constrained linear transformations (for certain features, such as cepstral features), we dedicate a separate chapter to them due to their slightly different motivation.

In a first attempt, Valbret et al. (1992b) proposed to warp the frequency axis based on pre-computed warping functions between source and target, using log-spectral features. The source speaker spectral tilt is subtracted before warping and the target speaker spectral tilt is added after warping. Some studies directly model and manipulate formant frequencies and bandwidths (Mizuno and Abe, 1995; Turajlic et al., 2003; Shuang et al., 2006; Godoy et al., 2010a) so that they match the target formants, as shown in Fig. 3. Maeda et al. (1999) proposed to cluster the acoustic space into different classes (similar to VQ) and perform a non-linear frequency warping on the STRAIGHT spectrum for each class. Later, Sündermann et al. (2003) studied various vocal tract length normalization (VTLN) approaches that were used in ASR to perform VC, including piecewise linear, power, quadratic, and bilinear VTLN functions. Erro et al. (2012) extended this VTLN approach to multiple classes and proposed an iterative algorithm to estimate the VTLN parameters. Přibilová and Přibil (2006) experimented with various linear and non-linear warping functions, with application to TTS adaptation. Erro and Moreno (2007b) proposed weighted frequency warping (WFW) to perform a piecewise linear frequency warping in each mixture components of a GMM, weighted by the posterior probability. It is worth noting that they used a pitch-asynchronous harmonic model (a high-quality vocoder) and performed phase manipulation to achieve high

quality speech. Toda et al. (2001) proposed to convert the source spectrum using a GMM and then warp the source spectrum to be similar to the converted spectrum with the aim of keeping the spectral details intact.

Other than the formant frequency locations, the average energy of the spectral bands is also an important factor in speaker individuality. Previously, this has been partly taken care of by subtracting source spectral tilt before frequency warping and adding the target spectral tilt. In an extension of WFW work, it was shown that in addition to frequency warping, an energy correction filter is required to increase the flexibility of the mapping function (Erro et al., 2010b). Tamura et al. (2011) proposed a simpler amplitude scaling by adding a *shift* value to the converted vector. In another extensive study, *amplitude scaling* in addition to frequency warping was proposed to add more degrees of freedom to the mapping (Godoy et al., 2011; Godoy et al.,2012).

Some frequency warping functions can be reformulated as a weighted linear mapping approach (Pitz and Ney, 2005). The linear mappings are usually constrained, so that the mapping is less prone to over-fitting. However, the constraints will limit the ability to mimic very different voices. Erro et al. (2013) studied this idea using bilinear warping function (similar to the VTLN approach) and constrained amplitude scaling, and extended it (Erro et al., 2015).

Numerous extensions of the FW approach have been proposed, such as in combination with GMMs (Erro et al., 2008; Zorilă et al., 2012; Mohammadi and Kain, 2013; Tian et al., 2015b), dictionary-based methods (Shuang et al., 2008; Uriz et al., 2009a; Tian et al., 2015a), and maximizing spectral correlation (Tian et al., 2014).

5.6. Adaptation techniques

In this section, we describe the techniques that are used when only limited or non-parallel training data are available. These approaches typically utilize the mappings or models learned from some pre-defined set of speakers to aid the training of the conversion mapping. Most of these approaches use mixture of linear mappings, however, the ideas could be generalized to other approaches such as neural networks.

Adaptation techniques perform mean adaptation on the means of GMM mixture components that are trained on the source speaker (Chen et al., 2003) to move the GMM means towards the target speaker's probability distribution. Mouchtaris et al. (2006) proposed an adaptation technique for non-parallel VC, in which a JDGMM is trained on a pre-defined set of source and target speakers that have parallel recordings. For building the mapping function using non-parallel recordings, the means and covariances of the GMMs are adapted to the new source and target speakers. In a neural network-based approach, a semi-supervised learning approach is proposed in which first speakers that sound similar to the source and target speakers are used for training the network, and then the pre-trained neural network is further trained using the source and target speaker data (Mohammadi and Kain, 2015). In another study, an adaptive RBM approach was proposed in which it is assumed that features are produced from a model where phonological information and speaker-related information are defined explicitly. During conversion, the phonetic and speaker information are separated and the speaker information is replaced with that of the target's (Nakashika et al., 2016).

Another scheme for voice conversion is to utilize the conversions built on multiple pre-stored speakers (different from the target speaker) to create the mapping function. A first attempt called speaker interpolation generates the target features using a weighted linear addition (interpolation) of multiple conversions towards multiple other pre-defined target speakers, by minimizing the difference between the target features and the converted features (Iwahashi and Sagisaka, 1994; 1995). The interpolation coefficients are estimated using only one word from the target speaker.

The eigenvoice VC (EVC) approach represents the joint probability density similar to the conventional GMM, except that the target means are defined as (Toda et al., 2006; Ohtani, 2010)

$$\mu_m^y = \mathbf{G}_m w + \mathbf{g}_m \tag{16}$$

where $\mathbf{g}_m$ is the bias vector and the matrix $\mathbf{G}_m = [\mathbf{g}_m^1, \ldots, \mathbf{g}_m^J]$ consists of basis vectors $\mathbf{g}_m^j$ for the mth mixture. The total number of basis vectors is J and the target speaker identity is controlled with the J-dimensional weight vector $w = [w^1, \ldots, w^J]^\top$. For a given target speaker, a weight is computed and assigned to each eigenvoice; the weight represents the eigenvoice's contribution to generating features (Toda et al., 2006; Ohtani, 2010). In the traditional eigenvoice approach, weights are estimated during training and are fixed during conversion. For lowering the computational cost, a multistep approach has been proposed (Masuda and Shozakai, 2007). For further improving the robustness of this approach to the amount of adaptation data, a maximum-a-posteriori adaptation approach has also been proposed (Tani et al., 2008). The eigenvoice approach has also been extended to *many-to-one* VC, where the target speaker is always the same but the source speaker can be an arbitrary speaker with minimal adaptation data (Toda et al., 2007b). Finally, *one-to-many* eigenvoice VC based on a tensor representation of the space of all speakers has been proposed (Saito et al., 2011). *Many-to-many* conversion has also been proposed in which the goal is to perform a conversion using an arbitrary source speaker to an arbitrary target speaker with minimal parallel (Ohtani et al., 2009) and non-parallel data (Ohtani et al., 2010).

5.7. Other mappings

Various other mapping approaches have been proposed. The *K*-histogram approach is a non-parametric approach which defines the mapping via the cumulative distribution function (CDF) of the source followed by an inverse CDF of the target (Uriz et al., 2009b)

$$\mathcal{F}_{\text{K-Histogram}}(\mathbf{x}) = CDF_y^{-1}(CDF_x(\mathbf{x})) \tag{17}$$

A Gaussian processes (GP) approach has also beens proposed (Pilkington et al., 2011; Xu et al., 2014). GPs are kernel-based, non-parametric approaches that can be viewed as distribution over functions, which relieves the need to specify the parametric form beforehand. For example, it is possible to define how to describe the mean and covariance functions (Pilkington et al., 2011). Another non-parametric approach based on topological maps has been proposed which estimates the joint distribution of the spectral space of source and target speakers (Rinscheid, 1996; Uchino et al., 2007). The topological map is a type of a neural network where each node is topologically located on a 2D map in a grid-like fashion. In the training step, the value of these nodes are learned. For each node in the source speaker map, a corresponding node in the target speaker map is computed. This correspondence is used to map an incoming source vector to a target vector. This approach has some similarities to the VQ method.

6. Prosodic modeling

Most of the VC literature focuses on mapping spectral features, despite the fact that prosodic aspects (pitch, duration, spectral balance, energy) are also important for speaker identity (Helander and Nurminen, 2007b; Morley et al., 2012). For modeling duration, a global speaking rate adjustment is not sufficient since it has been observed that phoneme durations differ somewhat arbitrarily between source and target speakers (Arslan and Talkin, 1998).

Table 3
An overview of pitch mapping methods for VC.

Method	Level	Pitch representation	Other info	Mapping function
Mean and variance matching (Chappell and Hansen, 1998)	Frame-level	F_0 contour	–	Linear
Predicting from spectrum (En-Najjary et al., 2003)	Frame-level	F_0 contour	Spectrum	Weighted linear
Joint modeling with spectrum (En-Najjary et al., 2004; Hanzlíček and Matoušek, 2007; Xie et al., 2014b)	Frame-level	F_0 contour	Spectrum	Weighted linear
Histogram equalization (Wu et al., 2010)	Frame-level	F_0 contour	–	Histogram equalization
MSD-HMM (Yutani et al., 2009)	Frame-level	F_0 contour	Spectrum	Weighted linear
LSTM (Chen et al., 2016; Ming et al., 2016)	Frame-level	F_0 contour	Spectrum	LSTM
Syllable-based codebook (Rao et al., 2007)	Syllable-level	F_0 contour	Syllable boundary	Codebook mapping
Syllable-based MLLR (Lolive et al., 2008)	Syllable-level	F_0 contour	Syllable boundary	MLLR adaptation
Syllable-based CART (Helander and Nurminen, 2007a)	Syllable-level	DCT	Syllable boundary	CART
Syllable-based weighted linear (Veaux and Rodet, 2011)	Syllable-level	DCT	Syllable boundary	Weighted linear
Hierarchical modeling of F0 (Sanchez et al., 2014)	Utterance-level	Wavelet transform (Suni et al., 2013)	–	KPLS (Helander et al., 2012)
Contour codebook + DTW (Chappell and Hansen, 1998; Inanoglu, 2003)	Utterance-level	F_0 contour	–	Codebook mapping
Weighting contours (Türk and Arslan, 2003; Inanoglu, 2003)	Utterance-level	F_0 contour	–	Weighting codebooks
SHLF parametrization (Gillett and King, 2003)	Utterance-level	Patterson (Patterson, 2000)	–	Piecewise linear
OSV parametrization (Ceyssens et al., 2002)	Utterance-level	Offset, slope and variance	–	Linear

Modeling duration using decision trees (Pozo, 2008) and duration-embedded HMMs has been studied (Wu et al., 2006).

The most common method to transform pitch is to globally match the average and standard deviation of the pitch contour. Pitch can be converted by mapping the log-scaled F_0 using a linear transformation

$$\hat{F}_0^y = \frac{\sigma^y}{\sigma^x}(F_0^x - \mu^x) + \mu^y \tag{18}$$

where μ and σ represent mean and standard deviation of the log-scaled F_0 (Chappell and Hansen, 1998). Several studies have looked into modeling F_0 and spectral features jointly (En-Najjary et al., 2004; Hanzlíček and Matoušek, 2007; Xie et al., 2014b); this has shown improvements for both spectral and F_0 conversions. Conversely, predicting pitch values from the target speaker spectrum using a GMM has also been studied (En-Najjary et al., 2003).

When we use simple linear mapping techniques, such as globally changing the speaking rate or adjusting the pitch mean and variance, the supra-segmental information is not modified effectively. Prosody modeling is a complex problem that depends on linguistic and semantic information. As an example, the emphasis that speakers put on certain speech units (such as words) does not necessarily have a similar pattern for other speakers depending on the context and high level information. In VC tasks, this high level information is typically not available. ASR can be used to automatically compute textual information, but the error that it is likely to introduce may become a detrimental factor for prosodic mapping performance. Pitch modeling for VC has been studied on different acoustic/linguistic levels: frame-level, syllable-level, and utterance-level. Moreover, various pitch representations have been used, such as F_0 contour, the discrete cosine transform (DCT) of the F_0 contour, the Wavelet transformation of the F_0 contour, and other compact parameterizations of the F_0 contour. In order to model the dynamics of the pitch contour in frame-level representations, mapping F_0 using multi-space probability distribution (MSD) HMMs (Yutani et al., 2009) and LSTM networks (Chen et al., 2016) have been proposed. Syllable-level representations model the pitch movements at the syllable level, which is a more meaningful representation for modeling pitch events. The most prominent pitch conversion approaches for VC are presented in Table 3. Wu et al. (2010) studied some of these approaches in more detail.

7. Performance evaluation

When evaluating the performance of VC systems several aspects can be evaluated:

Speaker similarity: Answers the question of "How similar is the converted speech to the target?". This is also known as conversion accuracy or speaker individuality.

Speech quality: This describes the quality of the generated speech with respect to naturalness and audible artifacts.

Speech intelligibility: Assesses the intelligibility of the generated speech. This is a lesser-studied aspect in the VC literature

In experimental voice conversion evaluations, a distinction is often made between intra-gender conversion (female-to-female or male-to-male) and inter-gender conversion (female-to-male or male-to-female).

A standard corpus for VC evaluation does not exist. Several databases have been used for VC including TIMIT (Garofolo et al., 1993), VOICES (Kain, 2001), CMU-Arctic (Kominek and Black, 2004), MOCHA (Wrench, 1999), and the MSRA mandarin corpus (Zhang et al., 2005). Very recently, the VC Challenge (VCC) 2016 prepared a standard dataset for a VC task, which has the potential to become the standard for VC studies (Toda et al., 2016). The VCtools available in the Festvox toolkit (Anumanchipalli et al., 2011) can be used for implementing baseline VC techniques such as GMM and MLPG/GV processing.

It has been shown that the performance of the system depends on the selection of source speaker. Turk and Arslan (2005) has studied the problem of automatic source speaker ("donor") selection from a set of available speakers that will result in the best quality output for a specific target speakers voice. This problem is also studied by proposing a selection measure (Hashimoto and Higuchi, 1995, 1996).

In the following subsections, we study the objective and subjective measures used for evaluating VC performance.

7.1. Objective evaluation

For evaluating VC performance objectively, a parallel-sentence corpus is required. First, the conversion and the associated target utterances have to be time-aligned. The difference between the converted speech and target can then be calculated using various general spectral difference measures. An example is the log-spectral distortion (in dB), which can be computed on unwarped, or warped (using the mel or Bark scale) spectra (Stylianou et al., 1998). The most prominent measure used in the VC literature is the mel-cepstrum distance (mel-CD), also measured in dB

$$\text{mel-CD}(\mathbf{y}, \hat{\mathbf{y}}) = (10/\ln 10)\sqrt{2(\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})} \tag{19}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are target and converted MCEP feature vectors, respectively.

The mel-CD is suitable for evaluating preliminary experiments, defining training criterions, and validation purposes, but not for evaluating the final system regarding quality due to the low correlation with human perception (Sündermann, 2005). Other objective speech quality assessment techniques exist (Rix et al., 2001). These measures aim to have higher correlation with human judgment. Recently, an automatic voice conversion evaluation strategy was proposed, wherein both speech quality and speaker similarity were automatically computed (Huang et al., 2016). The speaker similarity score was computed using a speaker verification method. These scores were shown to have higher correlation with subjective scores. However, optimizing mapping functions based on these criterions is more difficult, due to their the complex mathematical formulation.

7.2. Subjective evaluation

Unfortunately, objective evaluations do not necessarily correspond to human judgments. Thus, in most studies, subjective evaluations are performed; during such evaluations human listeners asses the performance of the VC system. The listeners should ideally perform their task in ideal listening environments; however, statistical requirements often necessitate a large number of listeners. Therefore, listeners are often hired that perform the task through a crowd-sourcing website such as Amazon Mechanical Turk (AMT).

The mean opinion score (MOS) test is an evaluation using 5-scale grading. Both the speech quality and similarity to the target voice can be evaluated. The grades are as follows: 5 = excellent, 4 = good, 3 = fair, 2 = poor, 1 = bad. The project TC-STAR proposes a standard perceptual MOS test as a measure of both quality and similarity (Sündermann et al., 2006b).

The comparative MOS (CMOS) can also be used to directly compare the speech quality of two VC techniques. The listener is asked to choose the better sounding utterance. The measure is computed as the percentage where each techniques is selected over the other. The grading can also be 5-scale as follows: 5 = definitely better, 4 = better, 3 = same, 2 = worse, 1 = definitely worse. This would give a good indication of any improvements. However, the absolute quality score is not calculated, making it difficult to judge the closeness to ideal quality (natural speech).

The ABX test is often used in comparing similarity between converted and target utterances. In this test, the listeners hears a pair of utterances A and B, followed by hearing a given utterance X, and have to decide is whether X is closer to A or B. The A and B utterances are uttered by source and target speakers but the ordering that the listener hears them is randomized. The measure is computed as the percentage of correct assignment of X to the target speaker. The main problem with interpreting ABX scores is that the subjects do not have the option to answer that the sentence X is not similar to neither A nor B (Machado and Queiroz, 2010). For example, given A = “mosquito”, B = “zebra”, X = “horse”, subjects may be forced to equate B with X; however, B is still very dissimilar from X.

The ABX test can compare two VC techniques directly by setting X, A, and B to the target utterance, first VC, and second VC. This measure is computed for each VC technique as the percentage of the utterances for which that technique has been chosen as closer to the target utterance. The MOS and ABX scores of various VC techniques have been published (Machado and Queiroz, 2010).

Another technique for testing similarity is to do use the CMOS for same-different testing (Kain, 2001). In this test, listeners hear two stimuli A and B with *different* content, and were then asked to indicate wether they thought that A and B were spoken by the *same*, or by two *different* speakers, using a five-point scale comprised of +2 (definitely same), +1 (probably same), 0 (unsure), - 1 (probably different), and - 2 (definitely different). One stimulus is the converted sample and the other is a *reference* speaker. Half of all stimuli pairs are created with the reference speaker identical to the target speaker of the conversion (the “same” condition); the other half were created with the reference speaker being of the same gender, but *not* identical to the target speaker of the conversion (the “different” condition). There has to be careful consideration in picking the proper speaker for the different condition.

Finally, the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test has been proposed to evaluate the speech quality of multiple stimuli. In this test, the subject is presented with a reference stimulus and multiple choices of test audio (stimuli), which they can listen to as many time as they want. The subjects are asked to score the stimuli according to a 5-scale score. This test is especially useful if one wants to test multiple system outputs in regards to speech quality.

As with all subjective testing, there is a lot of variability in the responses and it is highly recommended to perform proper significant testing on any subjective scores to show the reliability of improvements over baseline approaches. For crowd-sourcing experiments, it is best to incorporate certain sanity checks to exclude listeners that are performing below a minimum performance threshold, or inconsistently. A possible implementation of these recommendations is to include obviously good/bad stimuli in the experiment , and to duplicate a small percentage of trials.

An extensive subjective evaluation was performed during the 2016 VCC, with multiple submitted systems (Wester et al., 2016a). It was concluded that “there is still a lot of work to be done in voice conversion, it is not a solved problem. Achieving both high levels of naturalness and a high degree of similarity to a target speaker within one VC system remains a formidable task” (Wester et al., 2016a). The average quality MOS score was about 3.2 for top submissions. The similarity average score was around 70% correctly identified as target for top submissions. Due to the high number of entries, techniques to compare and visualize the high number of stimuli, such as multidimensional scaling, were utilized (Wester et al., 2016a, b).

8. Applications

VT and VC techniques can be applied to solve a variety of applications. We list some of these applications in this section:

- **Transforming speaker identity:** The typical application of VT is to transform speaker identity from one source speaker to a target speaker, which is referred to as VC (Childers et al., 1985). For example, a high-quality VC system could be used by dubbing actors to assume the original actor's voice characteristics. VT methods can also be applied for singing voice conversion (Turk et al., 2009; Villavicencio and Bonada, 2010; Doi et al., 2012; Kobayashi et al., 2013).
- **Transforming speaking type:** VT can be applied to transform the speaking type of a speaker. The goal is to retain the speaker identity but to transform emotion (Hsia et al., 2005; 2007; Tesser et al., 2010; Li et al., 2012), speaking style (Mohammadi et al., 2012; Godoy et al., 2013), speaker accent (Aryal et al., 2013), and speaker character (Pongkittiphan, 2012). Prosodic aspects are considered a more prominent factor in perceiving emotion and accent, thus some studies focus on prosodic aspects (Kawanami et al., 2003; Tao et al., 2006; Kang et al., 2006; Inanoglu and Young, 2007; Barra et al., 2007; Hsia et al., 2007; Li et al., 2012; Wang et al., 2012; Wang and Yu, 2014).

Personalizing TTS systems: A major application of VC is to personalize a TTS systems to new speakers, using limited amounts of training data from the desired speaker (typically the end-user if the TTS is used as an augmentative and alternative communications device) (Kain and Macon, 1998b; Duxans, 2006). Another option is to create a TTS system with new emotions (Kawanami et al., 2003; Türk and Schröder, 2008; Inanoglu and Young, 2009; Turk and Schroder, 2010; Latorre et al., 2014).

Speech-to-speech translation: The goal of these systems is to translate speech spoken in one language to another language, while preserving speaker identity (Wahlster, 2000; Bonafonte et al., 2006). These systems are usually a combination of ASR, followed by machine translation. Then, the translated sentence is synthesized using a TTS system in the destination language, followed by a cross-language VC system (Duxans et al., 2006; Sündermann et al., 2006b; Nurminen et al., 2006; Sündermann et al., 2006a).

Biometric voice authentication systems: VC presents a threat to speaker verification systems (Pellom and Hansen, 1999). Some studies have reported on the relation between the two systems and the vulnerabilities that VC poses for speaker verification, along with some solutions (Alegre et al., 2013; Wu et al., 2013b; Correia, 2014; Wu and Li, 2014).

Speaking- and hearing-aid devices: VT systems can potentially be used to help people with speech disorders by synthesizing more intelligible or more typical speech (Kain et al., 2007; Hironori et al., 2010; Toda et al., 2012b; Yamagishi et al., 2012; Aihara et al., 2013; Tanaka et al., 2013; Toda et al., 2014; Kain and Van Santen, 2009). VT is also applied in speaking-aid devices that use electrolarynx devices (Bi and Qi, 1997; Nakamura et al., 2006; 2012). Similar approaches can be used to increase the intelligibility of speech especially in noisy environments with application to increasing the performance of future hearing-aid devices (Mohammadi et al., 2012; Koutsogiannaki and Stylianou, 2014; Godoy et al., 2014). Other applications are devices that convert murmur to speech (Toda and Shikano, 2005; Nakagiri et al., 2006; Toda et al., 2012b), or whisper to speech (Morris and Clements, 2002; Tran et al., 2010).

Telecommunications: VT approaches have been used to reconstruct wide-band speech from its narrowband version (Park and Kim, 2000). This can enhance speech quality without modifying existing communication networks. Spectral conversion approaches have also been successfully used for speech enhancement (Mouchtaris et al., 2004b).

9. Challenges

Many unsolved problems exist in the area of VC. Some of them have been identified in previous studies (Childers et al., 1985; Kuwabara and Sagisak, 1995; Sündermann, 2005; Stylianou, 2009; Machado and Queiroz, 2010). As concluded in the VC Challenge 2016, there is still a significant performance gap between the current state-of-the-art performance levels and the human user expectations (Toda et al., 2016). There are a lot of similarities between components of VC and statistical TTS systems, since both aim to generate speech features and synthesizing waveforms (Ling et al., 2015). Consequently, some of the challenges and issues are shared in both systems.

Analysis/Synthesis issues: One major VC component that limits the quality of the generated speech is the analysis/synthesis part. STRAIGHT is a high-quality vocoder, but compared to natural speech, there is a still a quality gap (Kawahara et al., 2008). Recently, new high-quality vocoders were proposed, such as AHOCODER (Erro et al., 2011) and VOCAINE (Agiomyrgiannakis, 2015), both of which have shown improvements in statistical TTS. Recently, several first attempts for direct waveform modeling using neural networks for statistical parametric TTS were proposed (Tokuda and Zen, 2015; Kobayashi et al., 2015; Fan et al., 2015). These efforts may be a first step towards a new scheme for speech modeling/modification; however, the situation in VC is different since we have access to a valid source speaker utterance, which potentially allows copying certain aspects of speech without modifications.

Feature interpolation issues: To represent spectral envelopes, various features are used, such as spectral magnitude, all-pole representations (LSFs, LPCs), and cepstral features. One major issue with these features is that interpolating two spectral representations may not result in spectral representations that are generated by the human vocal tract. For example, when using cepstra, if we interpolate two different vowel regions, the outcome would sound as if the two sections are overlapping, and not as a single sound that lies perceptually between the two initial vowels. This limitation is one of the reasons for over-smoothing when multiple frames are averaged together. A spectral representation that represents meaningful features are formants locations and bandwidth. The two major problems of this representation is that formant extraction is still an unsolved problem, especially in noisy environments, and the inability of formants alone to represent finer spectral details.

One-to-many issues: The one-to-many problem in VC happens when two very similar speech segments of the source speaker have corresponding speech segments in the target speaker that are not similar to each other. As a result, the mapping function usually over-smoothes the generated features in order to be similar to both target speech segments. Some studies have attempted to solve this problem (Mouchtaris et al., 2007; Helander et al., 2008b).

Over-smoothing issues: In most VC approaches, the feature mapping is a result of averaging many parameters which results in over-smoothed features. This phenomenon is a symptom of the feature interpolation issue and one-to-many issue. This effect reduces both speech quality and speaker similarity. A lot of approaches such as GV have been proposed to increase the variability of the spectrum. Approaches like dictionary mapping and unit-selection don't suffer as much since they retain raw parameters and the feature manipulation is minimal; however, they typically require a larger training corpus and might suffer from discontinuous features and resulting audible discontinuities in the speech waveform.

Prosodic mapping issues: For converting prosodic aspects of speech, various methods have been proposed. However, most of them simply adjust some global statistics, such as average and standard deviation. The conversion is usually performed in the frame-level domain. As mentioned in the previous sections, these naive modifications can not effectively convert supra-segmental features. There are some challenges to modeling prosody for parametric VC. The main challenge is the absence of certain high-level features during conversion, which hugely affect human prosody. These features might be linguistic features (such as information about phonemes and syllables), or more abstract features (such as sarcasm and emotion). For TTS systems, textual information is available during conversion, which facilitates predicting prosodic features from more prosodically relevant representations such as syllable-level or word-level information. Es-

pecially foot-level information modeling might be helpful for conversion (Langarani and van Santen, 2015). These types of data, extracted from the input text, are not available to a stand-alone VC system, but could be extracted using ASR systems with some degree of error. The main challenge is to transform pitch contours by considering more context than one frame at a time, i. e. segmentally.

10. Future directions

In the previous section, we presented several challenges that current VC technology faces. In this section, we list some future research directions.

Non-Parallel VC: Most of the studies in the literature use parallel corpora. However, to make VC systems more mainstream, building transformation systems from non-parallel corpora is essential. The reason is that average users are hesitant to record numerous speech prompts with specific contents, which might be laborious for some. Several attempts for doing non-parallel VC is reported (Erro et al., 2010a; Nakashika et al., 2016).

Text-dependent VC: VC systems that utilize phonetic information are another research area. One example is to use phoneme identity before clustering the acoustic space (Kumar and Verma, 2003; Verma and Kumar, 2005). Using phonetic information to identify classes using a CART model instead of spectral information has also been proposed (Duxans et al., 2004). These systems could use the output of ASR to help the effectiveness of VC. These systems would likely use a combination of techniques from ASR, VC and parametric TTS.

Database size: An important research direction is capturing the voice using very limited recordings. Some studies propose methods for dealing with limited amounts of data (Hashimoto and Higuchi, 1996; Uto et al., 2006; Mesbahi et al., 2007b; Helander et al., 2008a; Popa et al., 2009; Tamura et al., 2011; Saito et al., 2012; Xu et al., 2014; Ghorbandoost et al., 2015). Utilizing additional unsupervised data have been proposed; for example, techniques that separate phonetic content and speaker identity are an elegant approach (Popa et al., 2009; Saito et al., 2012; Nakashika et al., 2016).

Modeling dynamics: Typically, most VC systems focus on performing transformations frame-by-frame. One approach to this consists of adding dynamic information to the mapping features. Event-based approaches seem to be a good representation since they decompose a sequence into events and transitions, and these can be individually modeled. However, detection of event locations is a challenging task and requires more research. Additionally, some models such as HMMs and RNNs implicitly model the speech dynamics from a sequence of local features. Typically, these models have higher number of parameters compared to frame-by-frame models. These sequence mapping approaches seem to be a major future direction.

Prosody modeling: Developing more complex prosody models that can capture speaker's intonation and segmental duration in an effective way is an important research direction. Most of the literature performs simple linear transformations of the pitch contour (typically in log domain) (Wu et al., 2010) and the speaking rate. Developing more sophisticated prosody models would enable the capture of complex prosodic patterns and thus enable more effective transformations.

Many-to-one conversion: In practice, most VC systems can only convert speech from the source speaker that they have been trained on. A more practical approach is to have a system that converts speech from anybody to the target speaker. Several attempts to accomplish this have been studied (Toda et al., 2007b).

Articulatory features: Most of the current literature studies the VC problem from a perceptual standpoint. However, it may be worthwhile to approach the problem from a speech production point of view. Several attempts to model and synthesize articulatory properties of the human vocal tract have been proposed (Toda et al., 2004; 2008). These approaches have some limitations, such as being speaker-dependent, or requiring hard-to-collect data such as MRI 3D images, electromagnetic articulography, and X-rays. Overcoming these limitations would open up an important set of tools for articulatory conversion and synthesis.

Perceptual optimization: The optimizations that are performed in statistical methods during learning source-target feature mapping function typically optimize criterions that are not highly correlated with human perception. An attempt at performing perceptual error optimization for DNN-based TTS has been proposed (Valentini-Botinhao et al., 2015); similar approaches could be adopted to VC.

Real-world situations: Most of the corpora used in the literature are recorded in clean conditions. In real-world situations, speech is often encountered in noisy environments. Attempts to perform VC on these noisy data would result in even more distorted synthesized speech. Creating corpora for these situations and developing noise-robust systems are an essential step to allowing VC systems to become mainstream.

References

Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., 1988. Voice conversion through vector quantization. In: Proceedings of the ICASSP.

Agiomyrgiannakis, Y., 2015. VOCAINE the vocoder and applications in speech synthesis. In: Proceedings of the ICASSP.

Agiomyrgiannakis, Y., Rosec, O., 2009. ARX-LF-based source-filter methods for voice modification and transformation. In: Proceedings of the ICASSP.

Aihara, R., Nakashika, T., Takiguchi, T., Ariki, Y., 2014a. Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary. In: Proceedings of the ICASSP.

Aihara, R., Takashima, R., Takiguchi, T., Ariki, Y., 2013. Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization. In: Proceedings of the ICASSP.

AIHARA, R., TAKIGUCHI, T., ARIKI, Y., 2015. Activity-mapping non-negative matrix factorization for exemplar-based voice conversion. In: Proceedings of the ICASSP.

Aihara, R., Takiguchi, T., Ariki, Y., 2015. Many-to-many voice conversion based on multiple non-negative matrix factorization. In: Proceedings of the INTERSPEECH.

Aihara, R., Ueda, R., Takiguchi, T., Ariki, Y., 2014b. Exemplar-based emotional voice conversion using non-negative matrix factorization. In: Proceedings of the APSIPA doi:10.1109/APSIPA.2014.7041640.

Alegre, F., Amehraye, A., Evans, N., 2013. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In: Proceedings of the ICASSP.

Anumanchipalli, G.K., Prahallad, K., Black, A.W., 2011. Festvox: Tools for creation and analyses of large speech corpora. Workshop on Very Large Scale Phonetics Research, UPenn, Philadelphia.

Arslan, L.M., 1999. Speaker transformation algorithm using segmental codebooks (STASC). Speech Commun. 28 (3), 211–226.

Arslan, L.M., Talkin, D., 1997. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In: Proceedings of the EUROSPEECH.

Arslan, L.M., Talkin, D., 1998. Speaker transformation using sentence HMM based alignments and detailed prosody modification. In: Proceedings of the ICASSP.

Aryal, S., Felps, D., Gutierrez-Osuna, R., 2013. Foreign accent conversion through voice morphing.. In: Proceedings of the INTERSPEECH.

Azarov, E., Vashkevich, M., Likhachov, D., Petrovsky, A., 2013. Real-time voice conversion using artificial neural networks with rectified linear units. In: Proceedings of the INTERSPEECH.

Barra, R., Montero, J.M., Macias-Guarasa, J., Gutiérrez-Arriola, J., Ferreiros, J., Pardo, J.M., 2007. On the limitations of voice conversion techniques in emotion identification tasks. In: Proceedings of the INTERSPEECH.

Benisty, H., Malah, D., 2011. Voice conversion using gmm with enhanced global variance.. In: Proceedings of the INTERSPEECH.
Benisty, H., Malah, D., Crammer, K., 2014. Sequential voice conversion using grid-based approximation. In: Proceedings of the IEEEI.
Bi, N., Qi, Y., 1997. Application of speech conversion to alaryngeal speech enhancement. IEEE Trans. Speech Audio Process. 5 (2), 97–105.
Bonafonte, A., Höge, H., Kiss, I., Moreno, A., Ziegenhain, U., van den Heuvel, H., Hain, H.-U., Wang, X.S., Garcia, M.-N., 2006. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In: Proceedings of the LREC.
Cano, P., Loscos, A., Bonada, J., De Boer, M., Serra, X., 2000. Voice morphing system for impersonating in karaoke applications. In: Proceedings of the ICMC.
Ceyssens, T., Verhelst, W., Wambacq, P., 2002. On the construction of a pitch conversion system. In: Proceedings of the EUSIPCO.
Chappell, D.T., Hansen, J.H., 1998. Speaker-specific pitch contour modeling and modification. In: Proceedings of the ICASSP.
Chen, L.-H., Ling, Z.-H., Dai, L.-R., 2014a. Voice conversion using generative trained deep neural networks with multiple frame spectral envelopes. In: Proceedings of the INTERSPEECH.
Chen, L.-H., Ling, Z.-H., Liu, L.-J., Dai, L.-R., 2014b. Voice conversion using deep neural networks with layer-wise generative training. IEEE/ACM Trans. Audio Speech Language Process. (TASLP) 22 (12), 1859–1872.
Chen, L.-H., Ling, Z.-H., Song, Y., Dai, L.-R., 2013. Joint spectral distribution modeling using restricted boltzmann machines for voice conversion. In: Proceedings of the INTERSPEECH.
Chen, L.-H., Liu, L.-J., Ling, Z.-H., Jiang, Y., Dai, L.-R., 2016. The USTC system for voice conversion challenge 2016: neural network based approaches for spectrum, aperiodicity and F0 conversion. In: Proceedings of the INTERSPEECH.
Chen, Y., Chu, M., Chang, E., Liu, J., Liu, R., 2003. Voice conversion with smoothed GMM and MAP adaptation. In: Proceedings of the EUROSPEECH.
Childers, D., Yegnanarayana, B., Wu, K., 1985. Voice conversion: Factors responsible for quality. In: Proceedings of the ICASSP.
Childers, D.G., 1995. Glottal source modeling for voice conversion. Speech Commun. 16 (2), 127–138.
Childers, D.G., Wu, K., Hicks, D., Yegnanarayana, B., 1989. Voice conversion. Speech Commun. 8 (2), 147–158.
Correia, M.J.R.F., 2014. Anti-Spoofing: Speaker Verification vs. Voice Conversion. Instituto Superior Técnico Master's Thesis.
Del Pozo, A., Young, S., 2008. The linear transformation of lf glottal waveforms for voice conversion.. In: Proceedings of the INTERSPEECH.
Desai, S., Black, A.W., Yegnanarayana, B., Prahallad, K., 2010. Spectral mapping using artificial neural networks for voice conversion. IEEE Trans. Audio Speech Lang. Process. 18 (5), 954–964.
Doi, H., Toda, T., Nakano, T., Goto, M., Nakamura, S., 2012. Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system. In: Proceedings of the APSIPA.
Dutoit, T., Holzapfel, A., Jottrand, M., Moinet, A., Perez, J., Stylianou, Y., 2007. Towards a voice conversion system based on frame selection. In: Proceedings of the ICASSP.
Duxans, H., 2006. Voice Conversion applied to Text-to-Speech systems. Universitat Politecnica de Catalunya, Barcelona, Spain Ph.D. thesis..
Duxans, H., Bonafonte, A., 2006. Residual conversion versus prediction on voice morphing systems. In: Proceedings of the ICASSP.
Duxans, H., Bonafonte, A., Kain, A., Van Santen, J., 2004. Including dynamic and phonetic information in voice conversion systems. In: Proceedings of the ICSLP.
Duxans, H., Erro, D., Pérez, J., Diego, F., Bonafonte, A., Moreno, A., 2006. Voice conversion of non-aligned data using unit selection. TC-STAR WSST.
En-Najjary, T., Rosec, O., Chonavel, T., 2003. A new method for pitch prediction from spectral envelope and its application in voice conversion. In: Proceedings of the INTERSPEECH.
En-Najjary, T., Rosec, O., Chonavel, T., 2004. A voice conversion method based on joint pitch and spectral envelope transformation.. In: Proceedings of the INTERSPEECH.
Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S., 2010. Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11, 625–660.
Erro, D., Alonso, A., Serrano, L., Navas, E., Hernáez, I., 2013. Towards physically interpretable parametric voice conversion functions. In: Advances in Nonlinear Speech Processing. Springer, pp. 75–82.
Erro, D., Alonso, A., Serrano, L., Navas, E., Hernaez, I., 2015. Interpretable parametric voice conversion functions based on gaussian mixture models and constrained transformations. Comput. Speech Lang. 30 (1), 3–15.
Erro, D., Alonso, A., Serrano, L., Tavarez, D., Odriozola, I., Sarasola, X., Del Blanco, E., Sanchez, J., Saratxaga, I., Navas, E., et al., 2016. Ml parameter generation with a reformulated mge training criterion—participation in the voice conversion challenge 2016. In: Proceedings of the INTERSPEECH.
Erro, D., Moreno, A., 2007a. Frame alignment method for cross-lingual voice conversion. In: Proceedings of the INTERSPEECH.
Erro, D., Moreno, A., 2007b. Weighted frequency warping for voice conversion.. In: Proceedings of the INTERSPEECH.
Erro, D., Moreno, A., Bonafonte, A., 2010a. INCA algorithm for training voice conversion systems from nonparallel corpora. IEEE Trans. Audio Speech Lang. Process. 18 (5), 944–953.
Erro, D., Moreno, A., Bonafonte, A., 2010b. Voice conversion based on weighted frequency warping. IEEE Trans. Audio Speech Lang. Process. 18 (5), 922–931.
Erro, D., Navas, E., Hernáez, I., 2012. Iterative MMSE estimation of vocal tract length normalization factors for voice transformation.. In: Proceedings of the INTERSPEECH.
Erro, D., Polyakova, T., Moreno, A., 2008. On combining statistical methods and frequency warping for high-quality voice conversion. In: Proceedings of the ICASSP.
Erro, D., Sainz, I., Navas, E., Hernáez, I., 2011. Improved HNM-based vocoder for statistical synthesizers.. In: Proceedings of the INTERSPEECH.
Eslami, M., Sheikhzadeh, H., Sayadiyan, A., 2011. Quality improvement of voice conversion systems based on trellis structured vector quantization. In: Twelfth Annual Conference of the International Speech Communication Association.
Fan, B., Lee, S.W., Tian, X., Xie, L., Dong, M., 2015. A waveform representation framework for high-quality statistical parametric speech synthesis. In: Proceedings of the APSIPA arXiv preprint arXiv:1510.01443.
Fujii, K., Okawa, J., Suigetsu, K., 2007. Highindividuality voice conversion based on concatenative speech synthesis. World Academy of Science, Engineering and Technology 2, 1.
Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustics, Speech and Signal Processing 34 (1), 52–59.
Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., 1993. DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. Nist Speech Disc 1-1.1, 93. NASA STI, Recon Technical Report N, p. 27403.
Ghorbandoost, M., Sayadiyan, A., Ahangar, M., Sheikhzadeh, H., Shahrebabaki, A.S., Amini, J., 2015. Voice conversion based on feature combination with limited training data. Speech Commun. 67, 113–128.
Gillett, B., King, S., 2003. Transforming f0 contours. In: Proceedings of the EUROSPEECH.
Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics, pp. 249–256.
Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks.. Aistats.
Godoy, E., Koutsogiannaki, M., Stylianou, Y., 2013. Assessing the intelligibility impact of vowel space expansion via clear speech-inspired frequency warping.. In: Proceedings of the INTERSPEECH.
Godoy, E., Koutsogiannaki, M., Stylianou, Y., 2014. Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles. Comput. Speech. Lang. 28 (2), 629–647.
Godoy, E., Rosec, O., Chonavel, T., 2009. Alleviating the one-to-many mapping problem in voice conversion with context-dependent modelling. In: Proceedings of the INTERSPEECH.
Godoy, E., Rosec, O., Chonavel, T., 2010a. On transforming spectral peaks in voice conversion. In: Proceedings of the SSW.
Godoy, E., Rosec, O., Chonavel, T., 2010b. Speech spectral envelope estimation through explicit control of peak evolution in time. In: Proceedings of the ISSPA.
Godoy, E., Rosec, O., Chonavel, T., 2011. Spectral envelope transformation using DFW and amplitude scaling for voice conversion with parallel or nonparallel corpora. Proceeding of the INTERSPEECH.
Godoy, E., Rosec, O., Chonavel, T., 2012. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. IEEE Trans. Audio Speech Lang. Process. 20 (4), 1313–1323.
Gu, H.-Y., Tsai, S.-F., 2014. Improving segmental GMM based voice conversion method with target frame selection. In: Proceedings of the ISCSLP.
Hanzlíček, Z., Matoušek, J., 2007. F0 transformation within the voice conversion framework. In: Proceedings of the INTERSPEECH.
Hashimoto, M., Higuchi, N., 1995. Spectral mapping method for voice conversion using speaker selection and vector field smoothing. In: Proceedings of the EUROSPEECH.
Hashimoto, M., Higuchi, N., 1996. Training data selection for voice conversion using speaker selection and vector field smoothing. In: Proceedings of the ICSLP.
Helander, E., Nurminen, J., Gabbouj, M., 2007. Analysis of lsf frame selection in voice conversion. In: Proceedings of the SPECOM.
Helander, E., Nurminen, J., Gabbouj, M., 2008a. Lsf mapping for voice conversion with very small training sets. In: Proceedings of the ICASSP.
Helander, E., Schwarz, J., Nurminen, J., Silen, H., Gabbouj, M., 2008b. On the impact of alignment on voice conversion performance. In: Proceedings of the INTERSPEECH.
Helander, E., Silén, H., Míguez, J., Gabbouj, M., 2010a. Maximum a posteriori voice conversion using sequential monte carlo methods.. In: Proceedings of the INTERSPEECH.
Helander, E., Silén, H., Virtanen, T., Gabbouj, M., 2012. Voice conversion using dynamic kernel partial least squares regression. IEEE Trans. Audio Speech Lang. Process. 20 (3), 806–817.
Helander, E., Virtanen, T., Nurminen, J., Gabbouj, M., 2010b. Voice conversion using partial least squares regression. IEEE Trans. Audio Speech Lang. Process. 18 (5), 912–921.
Helander, E.E., Nurminen, J., 2007a. A novel method for prosody prediction in voice conversion. In: Proceedings of the ICASSP.
Helander, E.E., Nurminen, J., 2007b. On the importance of pure prosody in the perception of speaker identity. In: Proceedings of the INTERSPEECH.
Hironori, D., Nakamura, K., Tomoki, T., Saruwatari, H., Shikano, K., 2010. Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models. IEICE Trans. Inf. Syst. 93 (9), 2472–2482.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2 (5), 359–366.
Hsia, C.-C., Wu, C.-H., Liu, T.-H., 2005. Duration-embedded bi-HMM for expressive voice conversion.. In: Proceedings of the INTERSPEECH.
Hsia, C.-C., Wu, C.-H., Wu, J.-Q., 2007. Conversion function clustering and selection using linguistic and spectral information for emotional voice conversion. IEEE Trans. Comput. 56 (9), 1245–1254.
Huang, D.-Y., Xie, L., Siu, Y., Lee, W., Wu, J., Ming, H., Tian, X., Zhang, S., Ding, C., Li, M., Nguyen, Q.H., Dong, M., Li, H., 2016. An automatic voice conversion evaluation strategy based on perceptual background noise distortion and speaker similarity. In: Proceedings of the SSW.
Hwang, H.-T., Tsao, Y., Wang, H.-M., Wang, Y.-R., Chen, S.-H., 2013. Incorporating global variance in the training phase of GMM-based voice conversion. In: Proceedings of the APSIPA.
Hwang, H.-T., Tsao, Y., Wang, H.-M., Wang, Y.-R., Chen, S.-H., et al., 2012. A study of mutual information for GMM-based spectral conversion.. In: Proceedings of the INTERSPEECH.
Imai, S., 1983. Cepstral analysis synthesis on the mel frequency scale. In: Proceedings of the ICASSP.
Imai, S., Kobayashi, T., Tokuda, K., Masuko, T., Koishida, K., Sako, S., Zen, H., 2009. Speech signal processing toolkit (SPTK), version 3.3.
Imai, S., Sumita, K., Furuichi, C., 1983. Mel log spectrum approximation (MLSA) filter for speech synthesis. Electron. Commun. Japan 66 (2), 10–18.
Inanoglu, Z., 2003. Transforming Pitch in a Voice Conversion Framework. St. Edmunds College, University of Cambridge Master's Thesis.
Inanoglu, Z., Young, S., 2007. A system for transforming the emotion in speech: combining data-driven conversion techniques for prosody and voice quality.. In: Proceedings of the INTERSPEECH, pp. 490–493.
Inanoglu, Z., Young, S., 2009. Data-driven emotion conversion in spoken english. Speech Commun. 51 (3), 268–283.
Iwahashi, N., Sagisaka, Y., 1994. Speech spectrum transformation by speaker interpolation. In: Proceedings of the ICASSP. Vol. 1. IEEE, pp. I–461.
Iwahashi, N., Sagisaka, Y., 1995. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. Speech Commun. 16 (2), 139–151.
Kain, A., Macon, M.W., 1998a. Spectral voice conversion for text-to-speech synthesis. In: Proceedings of the ICASSP.
Kain, A., Macon, M.W., 1998b. Text-to-speech voice adaptation from sparse training data.. In: Proceedings of the ICSLP.
Kain, A., Macon, M.W., 2001. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In: Proceedings of the ICASSP.
Kain, A., van Santen, J.P., 2007. Unit-selection text-to-speech synthesis using an asynchronous interpolation model.. In: Proceedings of the SSW.
Kain, A., Van Santen, J., 2009. Using speech transformation to increase speech intelligibility for the hearing-and speaking-impaired. In: Proceedings of the ICASSP.
Kain, A.B., 2001. High Resolution Voice Transformation. Oregon Health & Science University Ph.D. thesis.
Kain, A.B., Hosom, J.-P., Niu, X., van Santen, J.P., Fried-Oken, M., Staehely, J., 2007. Improving the intelligibility of dysarthric speech. Speech Commun. 49 (9), 743–759.
Kang, Y., Shuang, Z., Tao, J., Zhang, W., Xu, B., 2005. A hybrid gmm and codebook mapping method for spectral conversion. In: Affective Computing and Intelligent Interaction. Springer, pp. 303–310.
Kang, Y., Tao, J., Xu, B., 2006. Applying pitch target model to convert f0 contour for expressive mandarin speech synthesis. In: Proceedings of the ICASSP.
Kawahara, H., Masuda-Katsuse, I., De Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. Speech Commun. 27 (3), 187–207.
Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., 2008. TANDEM-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In: Proceedings of the ICASSP.
Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H., Shikano, K., 2003. GMM-based voice conversion applied to emotional speech synthesis. In: Proceedings of the EUROSPEECH.
Kim, E.-K., Lee, S., Oh, Y.-H., 1997. Hidden markov model based voice conversion using dynamic characteristics of speaker.. In: Proceedings of the EUROSPEECH.
Kobayashi, K., Doi, H., Toda, T., Nakano, T., Goto, M., Neubig, G., Sakti, S., Nakamura, S., 2013. An investigation of acoustic features for singing voice conversion based on perceptual age.. In: Proceedings of the INTERSPEECH.
Kobayashi, K., Toda, T., Neubig, G., Sakti, S., Nakamura, S., 2015. Statistical singing voice conversion based on direct waveform modification with global variance. In: Proceedings of the INTERSPEECH.
Kominek, J., Black, A.W., 2004. The CMU arctic speech databases. In: Proceedings of the SSW.
Koutsogiannaki, M., Stylianou, Y., 2014. Simple and artefact-free spectral modifications for enhancing the intelligibility of casual speech. In: Proceedings of the ICASSP.
Kumar, A., Verma, A., 2003. Using phone and diphone based acoustic models for voice conversion: a step towards creating voice fonts. In: Proceedings of the ICME.
Kuwabara, H., Sagisak, Y., 1995. Acoustic characteristics of speaker individuality: control and conversion. Speech Commun. 16 (2), 165–173.
Langarani, M.S.E., van Santen, J., 2015. Speaker intonation adaptation for transforming text-to-speech synthesis speaker identity. In: Proceedings of the ASRU.
Laskar, R., Chakrabarty, D., Talukdar, F., Rao, K.S., Banerjee, K., 2012. Comparing ANN and GMM in a voice conversion framework. Appl. Soft Comput. 12 (11), 3332–3342.
Laskar, R.H., Talukdar, F.A., Bhattacharjee, R., Das, S., 2009. Voice conversion by mapping the spectral and prosodic features using support vector machine. In: Applications of Soft Computing. Springer, pp. 519–528.
Latorre, J., Wan, V., Yanagisawa, K., 2014. Voice expression conversion with factorised HMM-TTS models. In: Proceedings of the INTERSPEECH.
Lee, C.-H., Wu, C.-H., 2006. Map-based adaptation for speech conversion using adaptation data selection and non-parallel training.. In: Proceedings of the INTERSPEECH.
Lee, K.-S., 2007. Statistical approach for voice personality transformation. IEEE Trans. Audio Speech Lang. Process. 15 (2), 641–651.
Lee, K.-S., 2014. A unit selection approach for voice transformation. Speech Commun. 60, 30–43.
Li, B., Xiao, Z., Shen, Y., Zhou, Q., Tao, Z., 2012. Emotional speech conversion based on spectrum-prosody dual transformation. In: Proceedings of the ICSP.
Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H.M., Deng, L., 2015. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. Signal Process. Mag. IEEE 32 (3), 35–52.
Liu, L.-J., Chen, L.-H., Ling, Z.-H., Dai, L.-R., 2014. Using bidirectional associative memories for joint spectral envelope modeling in voice conversion. In: Proceedings of the ICASSP.
Liu, L.-J., Chen, L.-H., Ling, Z.-H., Dai, L.-R., 2015. Spectral conversion using deep neural networks trained with multi-source speakers. In: Proceedings of the ICASSP.
Lolive, D., Barbot, N., Boeffard, O., 2008. Pitch and duration transformation with non-parallel data. In: Proceedings of the Speech Prosody.
Machado, A.F., Queiroz, M., 2010. Voice conversion: a critical survey. In: Proceedings of the SMC.
Maeda, N., Banno, H., Kajita, S., Takeda, K., Itakura, F., 1999. Speaker conversion through non-linear frequency warping of straight spectrum.. In: Proceedings of the EUROSPEECH.
Makki, B., Seyedsalehi, S., Sadati, N., Hosseini, M.N., 2007. Voice conversion using nonlinear principal component analysis. In: Proceedings of the CIISP.
Masaka, K., Aihara, R., Takiguchi, T., Ariki, Y., 2014. Multimodal voice conversion using non-negative matrix factorization in noisy environments. In: Proceedings of the ICASSP.
Masuda, T., Shozakai, M., 2007. Cost reduction of training mapping function based on multistep voice conversion. In: Proceedings of the ICASSP.
Matsumoto, H., Hiki, S., Sone, T., Nimura, T., 1973. Multidimensional representation of personal quality of vowels and its acoustical correlates. IEEE Trans. Audio Electroacoust. 21 (5), 428–436.
Matsumoto, H., Yamashita, Y., 1993. Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function.. J. Acoust. Soc. Japan (E) 14 (5), 353–361.
Mesbahi, L., Barreaud, V., Boeffard, O., 2007a. Comparing GMM-based speech transformation systems. In: Proceedings of the INTERSPEECH.
Mesbahi, L., Barreaud, V., Boeffard, O., 2007b. Gmm-based speech transformation systems under data reduction. In: Proceedings of the SSW.
Ming, H., Huang, D., Xie, L., Wu, J., Li, M.D.H., 2016. Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion. In: Proceedings of the INTERSPEECH.
Mizuno, H., Abe, M., 1995. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. Speech Commun. 16 (2), 153–164.
Mohammadi, S.H., Kain, A., 2013. Transmutative voice conversion. In: Proceedings of the ICASSP.
Mohammadi, S.H., Kain, A., 2014. Voice conversion using deep neural networks with speaker-independent pre-training. In: Proceedings of the SLT.
Mohammadi, S.H., Kain, A., 2015. Semi-supervised training of a voice conversion mapping function using a joint-autoencoder. In: Proceedings of the INTERSPEECH.
Mohammadi, S.H., Kain, A., 2016. A voice conversion mapping function based on a stacked joint-autoencoder. In: Proceedings of the INTERSPEECH.
Mohammadi, S.H., Kain, A., van Santen, J.P., 2012. Making conversational vowels more clear.. In: Proceedings of the INTERSPEECH.
Morise, M., 2015. Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. Speech Commun. 67, 1–7.
Morise, M., Yokomori, F., Ozawa, K., 2016. World: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. Syst..
Morley, E., Klabbers, E., van Santen, J.P., Kain, A., Mohammadi, S.H., 2012. Synthetic f0 can effectively convey speaker id in delexicalized speech.. In: Proceedings of the INTERSPEECH.
Morris, R.W., Clements, M.A., 2002. Reconstruction of speech from whispers. Med. Eng. Phys. 24 (7), 515–520.
Mouchtaris, A., Agiomyrgiannakis, Y., Stylianou, Y., 2007. Conditional vector quantization for voice conversion. In: Proceedings of the ICASSP.
Mouchtaris, A., Van der Spiegel, J., Mueller, P., 2004a. Non-parallel training for voice conversion by maximum likelihood constrained adaptation. In: Proceedings of the ICASSP.
Mouchtaris, A., Van der Spiegel, J., Mueller, P., 2004b. A spectral conversion approach to the iterative wiener filter for speech enhancement. In: Proceedings of the ICME.

Mouchtaris, A., Van der Spiegel, J., Mueller, P., 2006. Nonparallel training for voice conversion based on a parameter adaptation approach. IEEE Trans. Audio Speech Lang. Process. 14 (3), 952–963.
Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9 (5), 453–467.
Muramatsu, T., Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., 2008. Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory. In: Proceedings of the INTERSPEECH.
Nakagiri, M., Toda, T., Kashioka, H., Shikano, K., 2006. Improving body transmitted unvoiced speech with statistical voice conversion. In: Proceedings of the INTERSPEECH.
Nakamura, K., Toda, T., Saruwatari, H., Shikano, K., 2006. A speech communication aid system for total laryngectomies using voice conversion of body transmitted artificial speech. J. Acoust. Soc. Am. 120 (5), 3351.
Nakamura, K., Toda, T., Saruwatari, H., Shikano, K., 2012. Speaking-aid systems using GMM -based voice conversion for electrolaryngeal speech. Speech Commun. 54 (1), 134–146.
Nakashika, T., Takashima, R., Takiguchi, T., Ariki, Y., 2013. Voice conversion in high--order eigen space using deep belief nets. In: Proceedings of the INTERSPEECH.
Nakashika, T., Takiguchi, T., Ariki, Y., 2014a. High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion. In: Proceedings of the INTERSPEECH.
Nakashika, T., Takiguchi, T., Ariki, Y., 2015a. Sparse nonlinear representation for voice conversion. In: Proceedings of the ICME.
Nakashika, T., Takiguchi, T., Ariki, Y., 2015b. Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (3), 580–587. doi:10.1109/TASLP.2014.2379589.
Nakashika, T., Takiguchi, T., Ariki, Y., 2015c. Voice conversion using speaker-dependent conditional restricted Boltzmann machine. EURASIP J. Audio Speech Music Process. 2015 (1), 1–12.
Nakashika, T., Takiguchi, T., Minami, Y., 2016. Non-parallel training in voice conversion using an adaptive restricted boltzmann machine. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (11), 2032–2045.
Nakashika, T., Toru, Takiguchi, T., Tetsuya, Ariki, Y., Yasuo, 2014b. Voice conversion based on speaker-dependent restricted boltzmann machines. IEICE Trans. Inf. Syst. 97 (6), 1403–1410.
Nankaku, Y., Nakamura, K., Toda, T., Tokuda, K., 2007. Spectral conversion based on statistical models including time-sequence matching. In: Proceedings of the SSW.
Narendranath, M., Murthy, H.A., Rajendran, S., Yegnanarayana, B., 1995. Transformation of formants for voice conversion using artificial neural networks. Speech Commun. 16 (2), 207–216.
Nguyen, B.P., 2009. Studies on Spectral Modification in Voice Transformation. Japan Advanced Institute of Science and Technology Ph.D. thesis.
Nguyen, B.P., Akagi, M., 2007. Spectral modification for voice gender conversion using temporal decomposition. J.Signal Process.
Nguyen, B.P., Akagi, M., 2008. Phoneme-based spectral voice conversion using temporal decomposition and gaussian mixture model. In: Proceedings of the ICCE.
Nirmal, J., Patnaik, S., Zaveri, M.A., 2013. Voice transformation using radial basis function. In: Proceedings of the TITC. Springer, pp. 345–351.
Nirmal, J., Zaveri, M., Patnaik, S., Kachare, P., 2014. Voice conversion using general regression neural network. Appl. Soft Comput. 24, 1–12.
Nurminen, J., Popa, V., Tian, J., Tang, Y., Kiss, I., 2006. A parametric approach for voice conversion. In: TCSTAR WSST, pp. 225–229.
Nurminen, J., Tian, J., Popa, V., 2007. Voicing level control with application in voice conversion. In: Proceedings of the INTERSPEECH.
Ohtani, Y., 2010. Techniques for Improving Voice Conversion Based on Eigenvoices. Nara Institute of Science and Technology.
Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., 2006. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In: Proceedings of the INTERSPEECH.
Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., 2009. Many-to-many eigenvoice conversion with reference voice. In: Proceedings of the INTERSPEECH.
Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., 2010. Non-parallel training for many–to-many eigenvoice conversion. In: Proceedings of the ICASSP.
Paliwal, K.K., 1995. Interpolation properties of linear prediction parametric representations.. In: Proceedings of the EUROSPEECH.
Park, K.-Y., Kim, H.S., 2000. Narrowband to wideband conversion of speech using gmm based transformation. In: Proceedings of the ICASSP.
Patterson, D.J., 2000. linguistic Approach to Pitch Range Modelling. Edinburgh University Ph.D. thesis..
Pellom, B.L., Hansen, J.H., 1999. An experimental study of speaker verification sensitivity to computer voice-altered imposters. In: Proceedings of the ICASSP.
Percybrooks, W.S., Moore, E., 2008. Voice conversion with linear prediction residual estimaton. In: Proceedings of the ICASSP.
Pilkington, N.C., Zen, H., Gales, M.J., et al., 2011. Gaussian process experts for voice conversion. In: Proceedings of the INTERSPEECH.
Pitz, M., Ney, H., 2005. Vocal tract normalization equals linear transformation in cepstral space. Speech Audio Process. IEEE Trans. 13 (5), 930–944.
Pongkittiphan, T., 2012. Eigenvoice-Based Character Conversion and its Evaluations. The University of Tokyo Master's thesis..
Popa, V., Nurminen, J., Gabbouj, M., 2009. A novel technique for voice conversion based on style and content decomposition with bilinear models.. In: Proceedings of the INTERSPEECH.
Popa, V., Nurminen, J., Gabbouj, M., et al., 2011. A study of bilinear models in voice conversion. J. Signal Inf. Process. 2 (02), 125.
Popa, V., Silen, H., Nurminen, J., Gabbouj, M., 2012. Local linear transformation for voice conversion. In: Proceedings of the ICASSP.
Pozo, A., 2008. Voice Source and Duration Modelling for Voice Conversion and Speech Repair. University of Cambridge Ph.D. thesis..
Přibilová, A., Přibil, J., 2006. Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description. Speech Commun. 48 (12), 1691–1703.
Qiao, Y., Tong, T., Minematsu, N., 2011. A study on bag of gaussian model with application to voice conversion.. In: Proceedings of the INTERSPEECH, pp. 657–660.
Ramos, M.V., 2016. Voice Conversion with Deep Learning. Tecnico Lisboa Master's thesis..
Rao, K.S., Laskar, R., Koolagudi, S.G., 2007. Voice transformation by mapping the features at syllable level. In: Pattern Recognition and Machine Intelligence. Springer, pp. 479–486.
Rao, S.V., Shah, N.J., Patil, H.A., 2016. Novel pre-processing using outlier removal in voice conversion. In: Proceedings of the SSW.
Rentzos, D., Qin, S.V., Ho, C.-H., Turajlic, E., 2003. Probability models of formant parameters for voice conversion. In: Proceedings of the EUROSPEECH.
Rinscheid, A., 1996. Voice conversion based on topological feature maps and time–variant filtering. In: Proceedings of the ICSLP.
Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proceedings of the ICASSP.
Saito, D., Watanabe, S., Nakamura, A., Minematsu, N., 2012. Statistical voice conversion based on noisy channel model. IEEE Trans. Audio Speech Lang. Process. 20 (6), 1784–1794.
Saito, D., Yamamoto, K., Minematsu, N., Hirose, K., 2011. One-to-many voice conversion based on tensor representation of speaker space. In: Proceedings of the INTERSPEECH.
Salor, Ö., Demirekler, M., 2006. Dynamic programming approach to voice transformation. Speech communication 48 (10), 1262–1272.
Sanchez, G., Silen, H., Nurminen, J., Gabbouj, M., 2014. Hierarchical modeling of f0 contours for voice conversion. In: Proceedings of the INTERSPEECH.
Shikano, K., Nakamura, S., Abe, M., 1991. Speaker adaptation and voice conversion by codebook mapping. In: IEEE International Sympoisum on Circuits and Systems, pp. 594–597.
Shuang, Z., Bakis, R., Qin, Y., 2006. Voice conversion based on mapping formants. In: TC-STAR WSST, pp. 219–223.
Shuang, Z., Meng, F., Qin, Y., 2008. Voice conversion by combining frequency warping with unit selection. In: Proceedings of the ICASSP.
Shuang, Z.-W., Wang, Z.-X., Ling, Z.-H., Wang, R.-H., 2004. A novel voice conversion system based on codebook mapping with phoneme-tied weighting. In: Proceedings of the ICSLP.
Song, P., Bao, Y., Zhao, L., Zou, C., 2011. Voice conversion using support vector regression. Electron. Lett. 47 (18), 1045–1046.
Sorin, A., Shechtman, S., Pollet, V., 2011. Uniform speech parameterization for multi-form segment synthesis.. In: Proceedings of the INTERSPEECH.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Stylianou, I., 1996. Harmonic Plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification. Ecole Nationale Supérieure des Télécommunications Ph.D. thesis.
Stylianou, Y., 2009. Voice transformation: a survey. In: Proceedings of the ICASSP.
Stylianou, Y., Cappé, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. 6 (2), 131–142.
Sun, L., Kang, S., Li, K., Meng, H., 2015. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In: Proceedings of the ICASSP.
Sündermann, D., 2005. Voice conversion: State-of-the-art and future work. Fortschritte der Akustik 31 (2), 735.
Sündermann, D., 2008. Text-independent voice conversion. Universitätsbibliothek der Universität der Bundeswehr München Ph.D. thesis..
Sündermann, D., Bonafonte, A., Höge, H., Ney, H., 2004a. Voice conversion using exclusively unaligned training data. In: Proceedings of the ACL/SEPLN.
Sündermann, D., Bonafonte, A., Ney, H., Höge, H., 2004b. A first step towards text-independent voice conversion. In: Proceedings of the ICSLP.
Sündermann, D., Bonafonte, A., Ney, H., Höge, H., 2005. A study on residual prediction techniques for voice conversion.. In: Proceedings of the ICASSP.
Sündermann, D., Hoge, H., Bonafonte, A., Ney, H., Black, A., Narayanan, S., 2006a. Text-independent voice conversion based on unit selection. In: Proceedings of the ICASSP.
Sündermann, D., Höge, H., Bonafonte, A., Ney, H., Hirschberg, J., 2006b. TC-Star: cross-language voice conversion revisited. TC-Star Workshop. TC-Star Workshop.
Sündermann, D., Höge, H., Bonafonte, A., Ney, H., Hirschberg, J., 2006c. Text-independent cross-language voice conversion.. In: Proceedings of the INTERSPEECH.
Sündermann, D., Ney, H., 2003. An automatic segmentation and mapping approach for voice conversion parameter training. In: Proceedings of the AST.
Sündermann, D., Ney, H., Hoge, H., 2003. VTLN-based cross-language voice conversion. In: Proceedings of the ASRU.
Suni, A.S., Aalto, D., Raitio, T., Alku, P., Vainio, M., et al., 2013. Wavelets for intonation modeling in hmm speech synthesis. In: Proceedings of the SSW.
Takamichi, S., Toda, T., Black, A.W., Nakamura, S., 2014. Modulation spectrum-based post-filter for GMM-based voice conversion. In: Proceedings of the APSIPA.

Takamichi, S., Toda, T., Black, A.W., Nakamura, S., 2015. Modulation spectrum-constrained trajectory training algorithm for gmm-based voice conversion. In: Proceedings of the ICASSP.
Takashima, R., Aihara, R., Takiguchi, T., Ariki, Y., 2013. Noise-robust voice conversion based on spectral mapping on sparse space. In: Proceedings of the SSW.
Takashima, R., Takiguchi, T., Ariki, Y., 2012. Exemplar-based voice conversion in noisy environment. In: Proceedings of the SLT.
Tamura, M., Morita, M., Kagoshima, T., Akamine, M., 2011. One sentence voice adaptation using GMm-based frequency-warping and shift with a sub-band basis spectrum model. In: Proceedings of the ICASSP.
Tanaka, K., Toda, T., Neubig, G., Sakti, S., Nakamura, S., 2013. A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion.. In: Proceedings of the INTERSPEECH.
Tani, D., Toda, T., Ohtani, Y., Saruwatari, H., Shikano, K., 2008. Maximum a posteriori adaptation for many-to-one eigenvoice conversion. In: Proceedings of the INTERSPEECH.
Tao, J., Kang, Y., Li, A., 2006. Prosody conversion from neutral speech to emotional speech. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1145–1154.
Tao, J., Zhang, M., Nurminen, J., Tian, J., Wang, X., 2010. Supervisory data alignment for text-independent voice conversion. IEEE Trans. Audio Speech Lang. Process. 18 (5), 932–943.
Tesser, F., Zovato, E., Nicolao, M., Cosi, P., 2010. Two vocoder techniques for neutral to emotional timbre conversion.. In: Proceedings of the SSW.
Tian, X., Wu, Z., Lee, S., Chng, E.S., 2014. Correlation-based frequency warping for voice conversion. In: Proceedings of the ISCSLP. IEEE, pp. 211–215.
Tian, X., Wu, Z., Lee, S.W., Hy, N.Q., Chng, E.S., Dong, M., 2015a. Sparse representation for frequency warping based voice conversion. In: Proceedings of the ICASSP.
Tian, X., Wu, Z., Lee, S.W., Hy, N.Q., Dong, M., Chng, E.S., 2015b. System fusion for high-performance voice conversion. In: Proceedings of the INTERSPEECH.
Titterington, D.M., Smith, A.F., Makov, U.E., et al., 1985. Statistical Analysis of Finite Mixture Distributions, Vol. 7. Wiley New York.
Toda, T., Black, A.W., Tokuda, K., 2004. Acoustic-to-articulatory inversion mapping with gaussian mixture model.. In: Proceedings of the INTERSPEECH.
Toda, T., Black, A.W., Tokuda, K., 2005. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter.. In: Proceedings of the ICASSP.
Toda, T., Black, A.W., Tokuda, K., 2007a. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio Speech Lang. Process. 15 (8), 2222–2235.
Toda, T., Black, A.W., Tokuda, K., 2008. Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. Speech Commun. 50 (3), 215–227.
Toda, T., Muramatsu, T., Banno, H., 2012a. Implementation of computationally efficient real-time voice conversion.. In: Proceedings of the INTERSPEECH.
Toda, T., Nakagiri, M., Shikano, K., 2012b. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. IEEE Trans. Audio Speech Lang. Process. 20 (9), 2505–2517.
Toda, T., Nakamura, K., Saruwatari, H., Shikano, K., et al., 2014. Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. IEEE/ACM IEEE Trans. Audio Speech Lang. Process. 22 (1), 172–183.
Toda, T., Ohtani, Y., Shikano, K., 2006. Eigenvoice conversion based on gaussian mixture model. In: Proceedings of the INTERSPEECH.
Toda, T., Ohtani, Y., Shikano, K., 2007b. One-to-many and many-to-one voice conversion based on eigenvoices. In: Proceedings of the ICASSP.
Toda, T., Saito, D., Villavicencio, F., Yamagishi, J., Wester, M., Wu, Z., Chen, L.-H., et al., 2016. The voice conversion challenge 2016. In: Proceedings of the INTERSPEECH.
Toda, T., Saruwatari, H., Shikano, K., 2001. Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum. In: Proceedings of the ICASSP.
Toda, T., Shikano, K., 2005. NAM-to-speech conversion with gaussian mixture models. In: Proceedings of the INTERSPEECH.
Tokuda, K., Kobayashi, T., Imai, S., 1995. Speech parameter generation from HMM using dynamic features. In: Proceedings of the ICASSP.
Tokuda, K., Zen, H., 2015. Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis. In: Proceedings of the ICASSP.
Tran, V.-A., Bailly, G., Lœvenbruck, H., Toda, T., 2010. Improvement to a nam-captured whisper-to-speech system. Speech Commun. 52 (4), 314–326.
Turajlic, E., Rentzos, D., Vaseghi, S., Ho, C.-H., 2003. Evaluation of methods for parameteric formant transformation in voice conversion. Proceeding of the ICASSP.
Türk, O., 2007. Cross-Lingual Voice Conversion. Bogaziçi University Ph.D. thesis..
Türk, O., Arslan, L.M., 2003. Voice conversion methods for vocal tract and pitch contour modification.. In: Proceedings of the INTERSPEECH.
Turk, O., Arslan, L.M., 2005. Donor selection for voice conversion. In: Proceedings of the EUSIPCO.
Turk, O., Arslan, L.M., 2006. Robust processing techniques for voice conversion. Comput. Speech Lang. 20 (4), 441–467.
Turk, O., Buyuk, O., Haznedaroglu, A., Arslan, L.M., 2009. Application of voice conversion for cross-language rap singing transformation. In: Proceedings of the ICASSP.
Türk, O., Schröder, M., 2008. A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis.. In: Proceedings of the INTERSPEECH.
Turk, O., Schroder, M., 2010. Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. IEEE Trans. Audio Speech Lang. Process. 18 (5), 965–973.
Uchino, E., Yano, K., Azetsu, T., 2007. A self-organizing map with twin units capable of describing a nonlinear input–output relation applied to speech code vector mapping. Inf. Sci. 177 (21), 4634–4644.
Uriz, A., Aguero, P., Tulli, J., Gonzalez, E., Bonafonte, A., 2009a. Voice conversion using frame selection and warping functions. In: Proceedings of the RPIC.
Uriz, A., Agüero, P.D., Erro, D., Bonafonte, A., 2008. Voice Conversion Using Frame Selection. Reporte Interno Laboratorio de Comunicaciones-UNMdP
Uriz, A.J., Agüero, P.D., Bonafonte, A., Tulli, J.C., 2009b. Voice conversion using k-histograms and frame selection.. In: Proceedings of the INTERSPEECH.
Uto, Y., Nankaku, Y., Toda, T., Lee, A., Tokuda, K., 2006. Voice conversion based on mixtures of factor analyzers. Proceeding of the ICSLP.
Valbret, H., Moulines, E., Tubach, J.-P., 1992a. Voice transformation using PSOLA technique. In: Proceedings of the ICASSP.
Valbret, H., Moulines, E., Tubach, J.P., 1992b. Voice transformation using PSOLA technique. Speech Commun. 11 (2), 175–187.
Valentini-Botinhao, C., Wu, Z., King, S., 2015. Towards minimum perceptual error training for DNN-based speech synthesis. In: Proceedings of the INTERSPEECH.
Veaux, C., Rodet, X., 2011. Intonation conversion from neutral to expressive speech.. In: Proceedings of the INTERSPEECH.
Verma, A., Kumar, A., 2005. Voice fonts for individuality representation and transformation. ACM Trans. Speech Lang. Process. (TSLP) 2 (1), 4.
Villavicencio, F., Bonada, J., 2010. Applying voice conversion to concatenative singing-voice synthesis.. In: Proceedings of the INTERSPEECH.
Villavicencio, F., Bonada, J., Hisaminato, Y., 2015. Observation-model error compensation for enhanced spectral envelope transformation in voice conversion. In: Proceedings of the MLSP.
Vincent, D., Rosec, O., Chonavel, T., 2007. A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and HNM modeling. In: Proceedings of the ICASSP.
Wahlster, W., 2000. Verbmobil: Foundations of Speech-to-Speech Translation. Springer Science & Business Media.
Wang, M., Wen, M., Hirose, K., Minematsu, N., 2012. Emotional voice conversion for mandarin using tone nucleus model–small corpus and high efficiency. In: Proceedings of the Speech Prosody.
Wang, Z., Yu, Y., 2014. Multi-level prosody and spectrum conversion for emotional speech synthesis. In: Proceedings of the ICSP.
Watanabe, T., Murakami, T., Namba, M., Hoya, T., Ishida, Y., 2002. Transformation of spectral envelope for voice conversion based on radial basis function networks. In: Proceedings of the ICSLP.
Wester, M., Wu, Z., Yamagishi, J., 2016a. Analysis of the voice conversion challenge 2016 evaluation results. In: Proceedings of the INTERSPEECH.
Wester, M., Wu, Z., Yamagishi, J., 2016b. Multidimensional scaling of systems in the voice conversion challenge 2016. In: Proceedings of the SSW.
Wrench, A., 1999. The MOCHA-TIMIT articulatory database. Queen Margaret University College.
Wu, C.-H., Hsia, C.-C., Liu, T.-H., Wang, J.-F., 2006. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1109–1116.
Wu, Y.-C., Hwang, H.-T., Hsu, C.-C., Tsao, Y., Wang, H.-M., 2016. Locally linear embedding for exemplar-based spectral conversion. In: Proceedings of the INTERSPEECH.
Wu, Z., Chng, E.S., Li, H., 2013a. Conditional restricted boltzmann machine for voice conversion. In: Proceedings of the ChinaSIP.
Wu, Z., Chng, E.S., Li, H., 2014a. Joint nonnegative matrix factorization for exemplar-based voice conversion. In: Proceedings of the INTERSPEECH.
Wu, Z., Kinnunen, T., Chng, E., Li, H., 2010. Text-independent F0 transformation with non-parallel data for voice conversion.. In: Proceedings of the INTERSPEECH.
Wu, Z., Kinnunen, T., Chng, E.S., Li, H., 2012. Mixture of factor analyzers using priors from non-parallel speech for voice conversion. IEEE Signal Process. Lett. 19 (12), 914–917.
Wu, Z., Larcher, A., Lee, K.-A., Chng, E., Kinnunen, T., Li, H., 2013b. Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints.. In: Proceedings of the INTERSPEECH.
Wu, Z., Li, H., 2014. Voice conversion versus speaker verification: an overview. APSIPA Trans. Signal Inf. Process. 3, e17.
Wu, Z., Virtanen, T., Chng, E.S., Li, H., 2014b. Exemplar-based sparse representation with residual compensation for voice conversion. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 22 (10), 1506–1521.
Wu, Z., Virtanen, T., Kinnunen, T., Chng, E., Li, H., 2013c. Exemplar-based unit selection for voice conversion utilizing temporal information. In: Proceedings of the INTERSPEECH.
Wu, Z., Virtanen, T., Kinnunen, T., Chng, E.S., Li, H., 2013d. Exemplar-based voice conversion using non-negative spectrogram deconvolution. In: Proceedings of the SSW.
Xie, F.-L., Qian, Y., Fan, Y., Soong, F.K., Li, H., 2014a. Sequence error (se) minimization training of neural network for voice conversion. In: Proceedings of the INTERSPEECH.
Xie, F.-L., Qian, Y., Soong, F.K., Li, H., 2014b. Pitch transformation in neural network based voice conversion. In: Proceedings of the ISCSLP.
Xu, N., Tang, Y., Bao, J., Jiang, A., Liu, X., Yang, Z., 2014. Voice conversion based on gaussian processes by coherent and asymmetric training with limited training data. Speech Commun. 58, 124–138.

Yamagishi, J., Veaux, C., King, S., Renals, S., 2012. Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction. Acoust. Sci. Technol. 33 (1), 1–5.
Ye, H., Young, S., 2003. Perceptually weighted linear transformations for voice conversion.. In: Proceedings of the INTERSPEECH.
Ye, H., Young, S., 2004. Voice conversion for unknown speakers.. In: Proceedings of the INTERSPEECH.
Ye, H., Young, S., 2006. Quality-enhanced voice morphing using maximum likelihood transformations. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1301–1312.
Yue, Z., Zou, X., Jia, Y., Wang, H., 2008. Voice conversion using HMM combined with GMM. In: Proceedings of the CISP.
Yutani, K., Uto, Y., Nankaku, Y., Lee, A., Tokuda, K., 2009. Voice conversion based on simultaneous modelling of spectrum and f0. In: Proceedings of the ICASSP.
Zen, H., Nankaku, Y., Tokuda, K., 2011. Continuous stochastic feature mapping based on trajectory hmms. IEEE Trans. Audio Speech Lang. Process. 19 (2), 417–430.
Zhang, J., Sun, J., Dai, B., 2005. Voice conversion based on weighted least squares estimation criterion and residual prediction from pitch contour. In: Affective Computing and Intelligent Interaction. Springer, pp. 326–333.
Zhang, M., Tao, J., Nurminen, J., Tian, J., Wang, X., 2009. Phoneme cluster based state mapping for text-independent voice conversion. In: Proceedings of the ICASSP.
Zhang, M., Tao, J., Tian, J., Wang, X., 2008. Text-independent voice conversion based on state mapped codebook. In: Proceedings of the ICASSP.
Zolfaghari, P., Robinson, T., 1997. A formant vocoder based on mixtures of gaussians. In: Proceedings of the ICASSP.
Zorilă, T.-C., Erro, D., Hernaez, I., 2012. Improving the quality of standard GMM-based voice conversion systems by considering physically motivated linear transformations. In: Advances in Speech and Language Technologies for Iberian Languages. Springer, pp. 30–39.