TRANSMUTATIVE VOICE CONVERSION

Seyed Hamidreza Mohammadi and Alexander Kain

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

mohammah@ohsu.edu, kaina@ohsu.edu

ABSTRACT

There are two types of voice conversion (VC) systems: generative and transmutative. A generative VC system typically uses a compact parametrization of speech and *maps* input to output parameters directly; however, the relative low dimensionality of the underlying speech model reduces quality. On the other hand, a transmutative VC system *modifies* high-dimensional features of a high-fidelity speech model, leaving critical details unmodified. Two versions of transmutative VC approach are implemented and compared to a generative VC approach. The results show that the implemented transmutative VC is significantly better compared to generative VC in terms of quality. The difference between the two VC methods regarding recognition scores are insignificant.

Index Terms: voice conversion, speech transformation, frequency warping

1. INTRODUCTION

Speaker recognizability and speech quality are two important concerns for voice conversion systems, which digitally process a *source speaker*'s utterance to sound as if a *target speaker* had spoken it. Improvements in spectral quality are especially required for high-fidelity tasks such as movie dubbing or interpretive services. (Prosodic aspects of voice conversion are outside of the scope of this paper.) The level of quality for these types of tasks is typically not achievable with relatively compact, pole-zero models of speech, but instead necessitate high-dimensional models, such as those based on sine-waves [1]. However, these complex models have parameters of relatively high dimensionality, making it difficult to train a mapping function that predicts target parameters from source parameters, using relatively few training data.

An alternative approach to predicting high-dimensional output parameters directly from input parameters is to *modify* the high-dimensional input parameters, in effect using a more constrained mapping requiring fewer parameters. We will refer to the first method as *generative*, and to the second method as *transmutative*. In this paper, we propose two transmutative methods and evaluate them while also comparing to a generative method.

The paper is organized as follows: First, we formally introduce the key concepts of generative and transmutative voice conversion (Section 2). Then we detail the methods of the proposed voice conversion implementation (Section 3), followed by a perceptual evaluation by two listening tests (Section 4), before concluding (Section 5).

2. KEY CONCEPTS

Given parallel (same-content), concatenated utterance waveforms $src^{train}[t]$ and $trg^{train}[t]$, we first extract desired features $\tilde{X}_{N_s \times d}$ and $\tilde{Y}_{N_t \times d}$, where N_s and N_t represent the number of source and target frames, respectively, and d is the dimensionality of the feature vectors. Then, N frames of time-aligned features $X^{train}_{N\times d}$ and $Y^{train}_{N\times d}$ are constructed, using either dynamic time warping (DTW) [2] or knowledge of the phoneme boundaries, and subsequent interpolation in the feature domain. (Many approaches exist that overcome the requirement of parallel data [e. g. 3, 4], but without loss of generality we will use time-aligned features in this paper.) We will now contrast the generative with the transmutative approach.

2.1. Generative approach

Classically, during training, we find the optimal parameter set

$$\lambda^* = \arg\min_{\lambda} E\left(Y^{train}, \mathcal{F}(X^{train}, \lambda)\right) \tag{1}$$

where $F(\cdot, \lambda)$ is a feature mapping function with parameters λ, and E is an appropriate error function in the chosen feature domain. During conversion, we are given a new input source waveform $src^{test}[t]$ and its features X^{test} are mapped by evaluating

$$\mathcal{F}(X^{test}, \lambda^*) = \hat{Y}^{test}, \tag{2}$$

an approximation of the target features from which the final conversion waveform is computed.

Previous works have explored various implementation details. For example, types of speech features included formant frequencies [5], line spectral frequencies (LSF) [6], and cepstral features [7]. Types of mapping functions included vector quantization (VQ) [8], fuzzy-VQ [9], multivariate regression [10], artificial neural networks [5], regressive Gaussian mixture models [6, 11, 12], support vector machines [13], and trajectory models [14, 15].

2.2. Transmutative approach

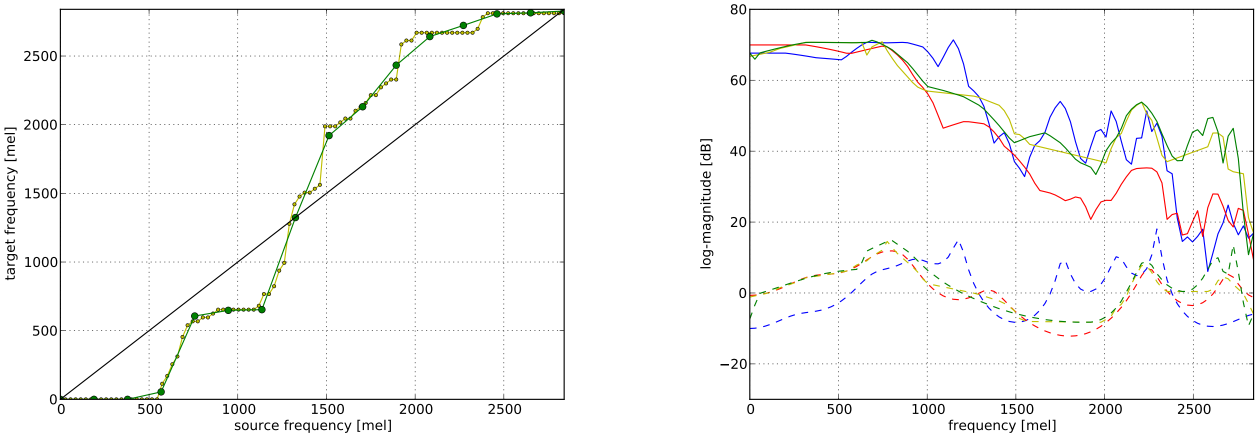
When the feature dimension d is large ($d > 100$), as is the case when using high-quality vocoders such as the harmonic model of speech, training becomes difficult due to the large number of parameters to be estimated, and the large amount of training data required. To address this problem, let us consider a system where we have high-dimensional features $X^{train}_{\uparrow}$ and $Y^{train}_{\uparrow}$. During training, we calculate the optimal parameter set

$$\lambda^*_{\mathcal{G}} = \arg\min_{\lambda_{\mathcal{G}}} E_{\mathcal{G}}\left(Y^{train}_{\uparrow}, \mathcal{G}(X^{train}_{\uparrow}, \lambda_{\mathcal{G}})\right) \tag{3}$$

where $\mathcal{G}(\cdot, \lambda)$ is an operation that *transmutes* its input $X^{train}_{\uparrow}$ according to parameters $\lambda_{\mathcal{G}}$, leaving critical details of its input unchanged. In other words, $\mathcal{G}$ is constrained in such a way that its possible outputs are congruent with the types of changes one expects *a priori* when converting one voice into another. Transmutation parameters $\lambda^*_{\mathcal{G}}$ can be predicted from a second, low-dimensional feature vector $X^{train}_{\downarrow}$ using a mapping function $\mathcal{H}(X^{train}_{\downarrow}, \lambda^*_{\mathcal{H}}) = \hat{\lambda}^*_{\mathcal{G}}$, with optimal parameters

$$\lambda^*_{\mathcal{H}} = \arg\min_{\lambda_{\mathcal{H}}} E_{\mathcal{H}}\left(\lambda^*_{\mathcal{G}}, \mathcal{H}(X^{train}_{\downarrow}, \lambda_{\mathcal{H}})\right). \tag{4}$$

This material is based upon work supported by the National Science Foundation under Grant No. 0964468.



(a) Frequency warping function (yellow) and its piece-wise linear parametrization (green). Green circles represent the "knots" or change-points of the piece-wise linear segmentation. The no-warp line (black) is added for reference.

(b) Source (blue) and target (red) magnitude spectra (solid lines), and their corresponding LPC spectra (blue and red dashed lines). Yellow lines are the result of applying the full (yellow) or parameterized warping function (green) to the source LPC (dashed) and original (solid) spectra.

Fig. 1: Example of spectral warping of one frame.

Finally, during conversion, we evaluate

$$\mathcal{G}\left(X_{\uparrow}^{test}, \mathcal{H}(X_{\downarrow}^{test}, \lambda_{\mathcal{H}}^{*})\right) = \hat{Y}_{\uparrow}^{test}. \qquad (5)$$

There are few examples of the transmutative approach in the literature. For example, Valbret et al. [10] were the first to modify the spectrum by dynamic frequency warping (DFW). Sundermann et al. and Erro et al. [16, 7] performed a constrained frequency warping similar to vocal tract length normalization. Finally, Erro et al. [17] used formant frequencies to calculate a more detailed warping function, and added a gain modification function. Godoy et al. [18] proposed amplitude scaling to modify gain.

3. TRANSMUTATIVE VOICE CONVERSION METHODS

The key idea behind the proposed transmutation algorithms is to let spectral warping and amplification take the role of $\mathcal{G}$, operating on high-dimensional sinusoidal parameters $X_{\uparrow}$ and $Y_{\uparrow}$. A probabilistic, piece-wise linear mapping function (with parameters $\lambda_{\mathcal{H}}$) takes the role of $\mathcal{H}$, predicting a parametrization $\lambda_{\mathcal{G}}$ of the warping and amplification functions. We detail two versions of the proposed algorithm, one deriving the warping function based on dynamic frequency warping of the linear predictive coded (LPC) spectrum (and using cepstral features as $X_{\downarrow}$), the other deriving the warping function based on formant-frequencies (and using them also as $X_{\downarrow}$).

3.1. DFW-LPC-based transmutation

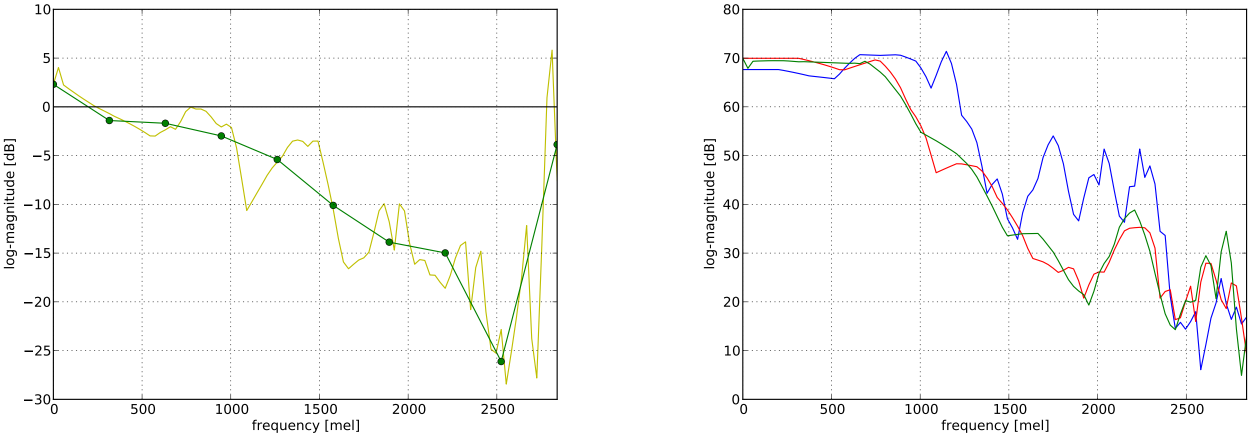
During training, we first computed pitch-synchronous spectral features based on manually corrected Glottal Closure Instances (GCIs) by encoding source and target waveforms using a harmonic vocoder [19]. We employed a Harmonic coder because of its high fidelity: a resynthesized waveform is typically perceptually indistinguishable from an unprocessed waveform. Harmonic analysis results in a F0 value and a variable number (dependent on F0) of harmonic amplitudes and phases per frame. The frame features were then time-aligned using either DTW or knowledge of phoneme boundaries, and nearest-neighbor interpolation. Then, for each frame, we resample the source and target magnitude spectra using 100 points on the Mel scale [20], using spline interpolation. Figure 1b shows corresponding example source and target magnitude spectra.

Second, we flattened both the source and the target magnitude spectra by removing their respective spectral tilts, and calculated a low-order LPC representation of them (see Figure 1b). We then used a DFW algorithm to compute the optimal path that aligns the LPC source and target magnitude spectra. An example warping function is shown in Figure 1a. Since we wish to predict this function, a compact parametrization of it was needed. We selected a piece-wise linear segmentation approach, calculated by optimizing a linear interpolation at 16 evenly-spaced frequencies (see Figure 1a), resulting in "knots" with globally constant x- and variable y-coordinates. We then applied the parametrized warping function to the magnitude spectrum and the unwrapped phase spectrum to calculate the warped spectrum, the results of which are shown in Figure 1b.

Third, we computed the spectral difference, referred to as the gain function, between the warped spectrum and the target spectrum and, analogously to the warping function, parametrized it using a piece-wise linear segmentation approach at 10 evenly spaced frequencies; an example of this is shown in Figure 2a. The warped and amplified spectrum is obtained by adding the parametrized gain function to the previously calculated warped magnitude spectrum; the result of this can be seen in Figure 2b.

Finally, we trained two independent regressive Joint-Density Gaussian Mixture Models (JDGMM) [6] with three mixture components (full covariance) each to predict the warping and gain function parameters from 16th-order cepstral features. Since both the warping and gain function parameters had constant x-coordinates of their knots, only y-coordinates needed to be predicted.

During conversion, we generated both the high-dimensional harmonic features and the low-dimensional cepstral features. Using the latter, the trained JDGMMs predicted transmutation parameters for both the warping and the gain functions, the results of which were applied to the harmonic features, and from which the final conversion waveform was computed. Figure 3 shows an excerpt of an example conversion. Pitch and duration changes were implemented within the harmonic vocoder framework.



(a) Gain function (yellow) and its piece-wise linear parametrization (green) using 10 "knots" (green circles). The zero-gain line (black) is added for reference.

(b) Source (blue), target (red), and warped and amplified source (green) magnitude spectra.

Fig. 2: Example of spectral amplification of one frame.

3.2. Formant-based transmutation

In the formant-based implementation, the spectral warping function was based on the first four source and target formant frequencies [similar to 17], obtained by automatic formant tracking [21, 22]. In this scenario, the x-coordinate of the warping knots is already known (the source formant frequencies), and the y-coordinates (corresponding to the target formant frequencies) of knots had to be predicted from the source formant frequencies, resulting in a compact 4→4-dimensional mapping.

The final warping function was constructed by subtracting and adding a constant bandwidth of 150 Hz to both the x- and the y-coordinate of each of the four knots, while specially handling situations where these frequencies overlapped; thus, the final warping function has 10 knots (adding required knots at zero and one at a maximum frequency). This approach ensured that regions around a formant, which carry important bandwidth information, were not adversely affected by a warping function that does not have a slope close to one in that region.

Since this implementation already required knowledge of formant frequencies, we also used source formants frequencies to predict y-coordinates of the gain knots required to reconstruct the gain function, obviating the need for cepstral features.

4. EVALUATION

We used 70 Harvard sentences [23] spoken by two male (M1, M2) and two female speakers (F1, F2) as speech material (sampling rate of 16 kHz). We restricted the set of possible conversion to two cross-gender (M1→F1, F2→M2) and two intra-gender (M2→M1, F1→F2) conversions, for a total of four different conversions. We used 46 of the sentences to train the two variations of transmutative conversion methods and one generative system; 4 sentences were used as a development set. For the generative system, we trained a JDGMM to map 18th-order LSF source parameters to same-order LSF target parameters, in an impulse/noise-excited LPC analysis/synthesis framework. Since its introduction, the JDGMM mapping paradigm has had many extensions and refinements [e. g. 12], but since we used the same basic methods for all three systems, their performance can be viewed as relative. All three conversion systems incorporated a basic prosody conversion approach, which included modifying the source's F0 mean and variance, as well as the average speaking rate to match the target's prosody statistics.

We created the following five stimulus conditions from the remaining 20 test sentences: natural waveform (NAT), DFW-LPC-based transmutative conversion (DFW), formant-based transmutative conversion (FOR), LSF-based generative conversion (GEN), and LSF vocoder resynthesis (LSF). After creation, all stimuli were loudness-normalized using an A-weighted [24] RMS measure.

4.1. Speaker recognition test

To evaluate conversion performance, we conducted a same-difference speaker recognition test [25]. In this test, listeners hear two utterances A and B with *different* content, and are 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 consisting of +2 (definitely same), +1 (probably same), 0 (unsure), -1 (probably different), and -2 (definitely different). We considered the following five stimulus pairs: NAT-NAT, NAT-DFW, NAT-FOR, NAT-GEN, and NAT-LSF. The first pair measured human performance as a reference, the middle pairs measured conversion performance, and the last pair measured the degradation due to using a compact parametric vocoder.

The listening experiment was carried out using Amazon Mechanical Turk [26]. The experiment was administered to 44 listeners, all of whom had approval ratings of at least 90%. Each listener judged 40 sentence pairs, 10 trials for each of the four conversions. During these 10 trials, 2 trials were used for each of the 5 conditions. Each test sentence was used exactly four times. The presentation order was randomized, but the sentence pair assignments remained fixed. Half of the pairs involved the "same" speaker (the conversion and the target), and the other half involved the source speaker for intra-gender conversions, and the alternate speaker of the same gender for cross-gender conversion.

Results are presented in Table 1. A two-tail t-test show an statistically insignificant difference between the NAT-FOR and NAT-GEN ($p = 0.24$). The interesting result is the massive difference between NAT-FOR and NAT-GEN in "same" and "diff" conditions. In the

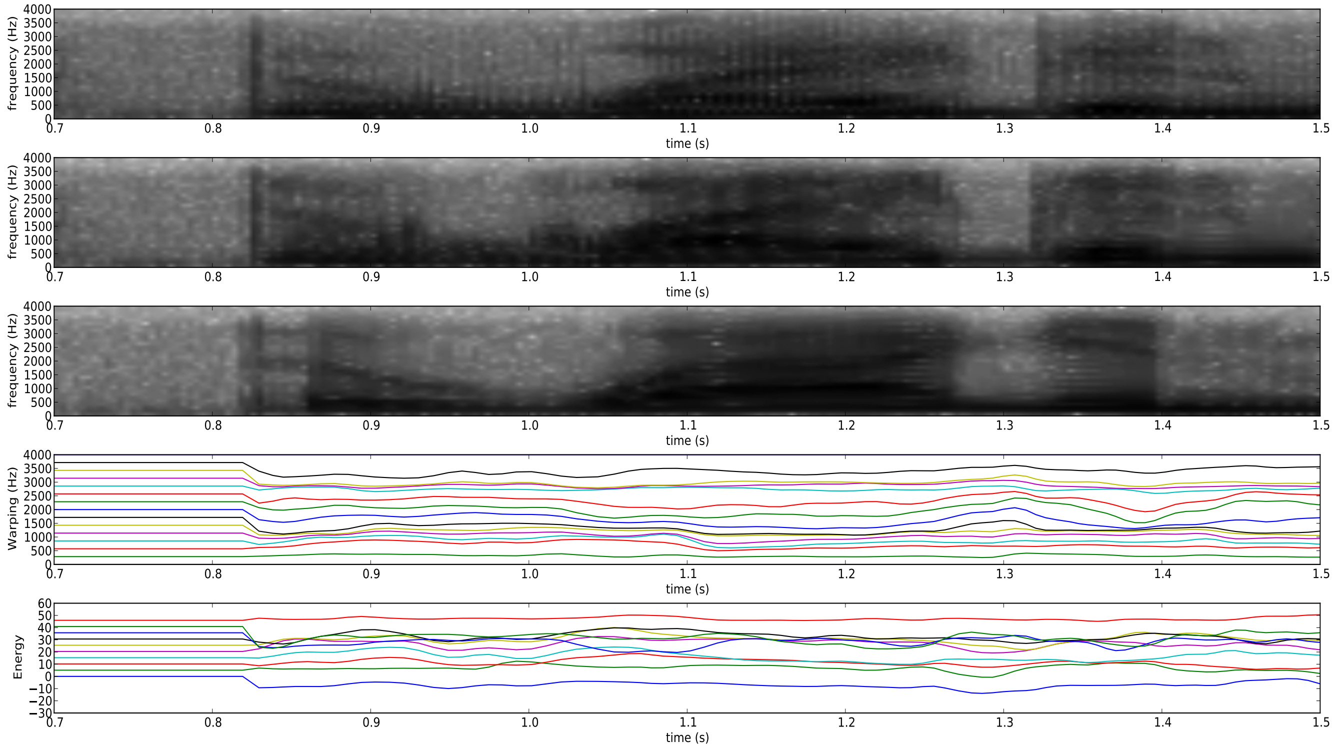


Fig. 3: Source (first panel), conversion (second panel), and target (third panel, time-aligned to the source for comparison purposes only) spectrograms, as well as corresponding warping (fourth panel) and gain (added by an arbitrary value for visualization) (fifth panel) parameter trajectories for the LPC-based conversion, for the utterance "mesh wire".

NAT-	NAT	DFW	FOR	GEN	LSF
same	1.39	-0.37	-0.38	0.12	1.04
diff	-1.32	-0.29	-0.68	-0.22	-1.08
all	1.36	-0.039	0.14	0.17	1.06

Table 1: Average speaker recognition test results (standard deviation is between 1.0 to 1.2) for diff, same and all conditions.

"diff" condition, FOR gets a better score, compared to the "same" condition where FOR result is far worse than GMM. It seems that because of the lower degree of freedom in DFW and FOR (only frequency warping and gain modification), it is not very capable of having various kinds of mappings.

4.2. Quality preference test

To evaluate conversion speech quality, we conducted a comparative mean opinion score (CMOS) test. In this test, listeners hear two utterances A and B with the *same* content and the *same* speaker but two *different* conditions, and are then asked to indicate wether they thought B was better or worse than A, using a five-point scale consisting of +2 (much better), +1 (somewhat better), 0 (same), -1 (somewhat worse), -2 (much worse). We considered the following four stimulus pairs: FOR-GEN, DFW-GEN, GEN-NAT, FOR-NAT. The first two pairs compare conversions directly against each other, and the last two measure conversion quality compared to natural waveforms.

The preference test was carried out identically to the speaker recognition test, except each test sentence was used exactly two times, and the order of A and B were randomized for each trial. A total of 35 listeners evaluated the results.

	FOR-GEN	DFW-GEN	GEN-NAT	FOR-NAT
all	-0.43(1.4)	0.88(0.9)	1.83(0.4)	1.57(1.1)

Table 2: Average preference test results (standard deviation in parentheses).

Results are presented in Table 2. The results indicate the superiority of the FOR compared to GEN in two ways. First is the direct comparison in which FOR is picked more. Also compared to NAT, FOR is picked more often than GEN. DFW method is picked less because of the audible jumps during parameter estimation which results in a lower quality. The two-tail t-test shows a statistically significance difference between FOR-NAT and GEN-NAT ($t(188) = 2.24$, $p = 0.026$).

5. CONCLUSION

VC methods can be categorized in two groups: generative and transmutative. Generative methods try to map in a compact parameters space. A synthesis filter is then used to synthesize the output. Transmutative methods try to modify high-dimensional parameters to keep the fine details of the spectrum. The results showed that transmutative methods have significantly higher quality scores with almost the same recognition score. The difference between recognition results of NAT-FOR in "diff" and "same" conditions showed that FOR is not changing the speaker identity very much. This may be because the current transmutative method has limited modification possibilities. Adding to the degree of freedom may possibly improve the recognition score.

6. REFERENCES

[1] Thomas F. Quatieri, *Discrete-time speech processing: principles and practice*, Prentice-Hall, 2002.

[2] H. Sakoe and S. Chiba, “A dynamic programming approach to continuous speech recognition,” in *Proceedings of the Seventh International Congress on Acoustics*, 1971, vol. 3, pp. 65–69.

[3] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, “Nonparallel training for voice conversion by maximum likelihood constrained adaptation,” in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP’04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–1.

[4] D. Erro, A. Moreno, and A. Bonafonte, “INCA algorithm for training voice conversion systems from nonparallel corpora,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 944–953, 2010.

[5] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, “Transformation of formants for voice conversion using artificial neural networks,” *Speech Communication*, vol. 16, no. 2, pp. 207–216, 1995.

[6] A. Kain and M. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *Proceedings of ICASSP*, May 1998, vol. 1, pp. 285–299.

[7] D. Erro, E. Navas, and I. Hernaez, “Iterative MMSE estimation of vocal tract length normalization factors for voice transformation,” in *Proc. Interspeech*, 2012.

[8] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” in *Proceedings of ICASSP*, 1988, pp. 655–658.

[9] K. Shikano, S. Nakamura, and M. Abe, “Speaker adaptation and voice conversion by codebook mapping,” in *Circuits and Systems, 1991., IEEE International Sympoisum on*. IEEE, 1991, pp. 594–597.

[10] H. Valbret, E. Moulines, and J. P. Tubach, “Voice transformation using PSOLA technique,” *Speech Communication*, vol. 11, no. 2, pp. 175–187, 1992.

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

[12] T. Toda, A. W. Black, and Tokuda K., “Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter,” in *Proceedings of ICASSP*, 2005, pp. 9–12.

[13] R. Laskar, F. Talukdar, R. Bhattacharjee, and S. Das, “Voice conversion by mapping the spectral and prosodic features using support vector machine,” *Applications of Soft Computing*, pp. 519–528, 2009.

[14] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Transactions on Audio, Speech, and Language Processing Journal*, vol. 15, no. 8, pp. 2222–2235, November 2007.

[15] H. Zen, Y. Nankaku, and K. Tokuda, “Continuous stochastic feature mapping based on trajectory HMMs,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 417–430, 2011.

[16] D. Sundermann, H. Ney, and H. Hoge, “VTLN-based cross-language voice conversion,” in *Automatic Speech Recognition and Understanding, 2003. ASRU’03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 676–681.

[17] D. Erro and A. Moreno, “Weighted frequency warping for voice conversion,” in *Proc. Interspeech*, 2007, pp. 1965–1968.

[18] E. Godoy, O. Rosec, and T. Chonavel, “Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1313–1323, May 2012.

[19] Y. Stylianou, J. Laroche, and E. Moulines, “High-quality speech modification based on a harmonic+noise model,” in *Proceedings of EUROSPEECH*, 1995, vol. 1, pp. 451–454.

[20] S.S. Stevens, J. Volkmann, and EB Newman, “A scale for the measurement of the psychological magnitude pitch,” *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[21] David Talkin, “Speech formant trajectory estimation using dynamic programming with modulated transition costs,” *The Journal of the Acoustical Society of America*, vol. 82, no. S1, pp. S55–S55, 1987.

[22] K. Sjölander and J. Beskow, “WaveSurfer — an open source speech tool,” in *Proc. of ICSLP*, 2000, pp. 464–467.

[23] J.P. Egan, “Articulation testing methods,” *The Laryngoscope*, vol. 58, no. 9, pp. 955–991, 1948.

[24] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, Acad. Press, 2003.

[25] A. Kain, *High Resolution Voice Transformation*, Ph.D. thesis, OGI School of Science & Engineering at Oregon Health & Science University, 2001.

[26] M. Buhrmester, T. Kwang, and S. D. Gosling, “Amazon’s mechanical turk — a new source of inexpensive, yet high-quality, data?,” *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, January 2011.