Vulnerability of Speaker Verification Systems Against Voice Conversion Spoofing Attacks: the Case of Telephone Speech

Tomi Kinnunen[1], *Zhi-Zheng Wu*[2], *Kong Aik Lee*[3], *Filip Sedlak*[1], *Eng Siong Chng*[2], *Haizhou Li*[3]

[1]School of Computing, University of Eastern Finland (UEF), Joensuu, Finland
[2]School of Computer Engineering, Nanyang Technological University (NTU), Singapore
[3]Human Language Technology Department, Institute for Infocomm Research (I^2R), Singapore
tkinnu@cs.joensuu.fi, wuzz@ntu.edu.sg, kalee@i2r.a-star.edu.sg

Abstract

Voice conversion – the methodology of automatically converting one's utterances to sound as if spoken by another speaker – presents a threat for applications relying on speaker verification. We study vulnerability of text-independent speaker verification systems against voice conversion attacks using telephone speech. We implemented a voice conversion systems with two types of features and nonparallel frame alignment methods and five speaker verification systems ranging from simple Gaussian mixture models (GMMs) to state-of-the-art joint factor analysis (JFA) recognizer. Experiments on a subset of NIST 2006 SRE corpus indicate that the JFA method is most resilient against conversion attacks. But even it experiences more than 5-fold increase in the false acceptance rate from 3.24 % to 17.33 %.

Index Terms: speaker verification, voice conversion, security

1. Introduction

Speaker verification is the task of accepting or rejecting an identity claim based on a speech sample [1]. Although recognition accuracy of speaker verification systems has considerably increased in the past few years thanks to intersession variability compensation techniques (e.g. [2]), in practice few people would trust a security system, such as e-banking application, relying solely on speaker verification. A common argument is that an intruder may use simple spoofing techniques to act as another speaker - the most obvious would be playback of an earlier recorded target speaker's voice. To respond to such concerns, a number of authors have studied how speaker recognition systems respond to playback attacks [3, 4], speaker-adapted speech synthesis [5, 6, 7], voice conversion [8, 9] and even human voice mimicking [10, 11]. While the datasets, spoofing techniques and recognition systems are rather diverse in these studies, they all clearly indicate significantly increased false acceptance rates under spoofing attacks. The concern about security of speaker verification, therefore, is well justified.

In this study, we apply *voice conversion* [12] techniques to simulate spoofing attacks (Fig. 1). Voice conversion techniques modify one speaker's (the source) utterances so that they sound as spoken by another speaker (the target). Voice conversion systems consist of training and conversion phases. In both phases, speech signal is first parameterized into short-term feature vectors. In the training phase, source and target speaker features are first paired at frame-level, typically using parallel training utterances. A stochastic Gaussian mixture model (GMM) conversion function is then trained using the paired vectors [13, 14, 12]. In the conversion phase, the conversion function is used for mapping unseen source features towards the target speaker. The converted utterance is reconstructed using inverse parameterization.

Previous studies on spoofing attacks have mostly considered high-quality speech recordings, relatively small number of speakers and typically just one speaker verification systems. Due to great potential of speaker verification in remote authentication tasks over non-ideal transmission channels, we would like to take the challenge to verify (or nullify) whether voice conversion spoofing poses a real threat on telephone speech. To this end, we pick the NIST 2006 SRE corpus for our experiments. Converting telephony speech poses practical challenges due to lacking parallel training corpus and low-quality signals with transmission channel effects.

The authors of [9] studied vulnerability of GMM recognizer against voice conversion attacks also on the SRE 2006 corpus. In the present study we carry out more thorough comparison including five speaker recognizers. Three of these [15, 16, 17] – used for reference purposes – are lightweight recognizers without intersession compensation or external score normalizations. The other two, GMM supervector [18] with nuisance attribute projection (NAP) [19] and state-of-the-art GMM with joint factor analysis (JFA) [2], in turn, include intersession compensation and score normalization. Even though the latter two can handle challenging cross-channel conditions very well, it is less obvious how they would respond to test utterances processed through voice conversion; their speaker models, background models, session and session variability models and score normalization cohort models are all trained using natural speech. Preliminary evaluation of JFA robustness against four types of spoofing and tampering attacks was studied in [4] using a small set of nonpublic data. The present study includes large number of data and larger pool of recognition systems.

2. Designing the Corpus

Due to prevalence of telephones and potential of speaker recognition technology in remote access applications, we decided to focus on telephony speech. To this end, we choose a subset of the core task in the NIST 2006 SRE corpus[1] as our **baseline corpus**. Our target speaker model training utterances and the verification trials are directly taken as a subset of the 1conv4w-1conv4w task in the original corpus. Our speaker detection task

The work of T. Kinnunen was supported by Academy of Finland (project no. 132129). Computer resources from *CSC - IT Center for Science* were used for part of the recognition experiments.

[1] http://www.itl.nist.gov/iad/mig//tests/sre/2006/index.html

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

Table 1: Statistics of the trials (subset of NIST SRE 2006 core).

	Female	Male	Total
Target speakers	298	206	504
Genuine trials	2349	1629	3978
Impostor trials	1636	1146	2782

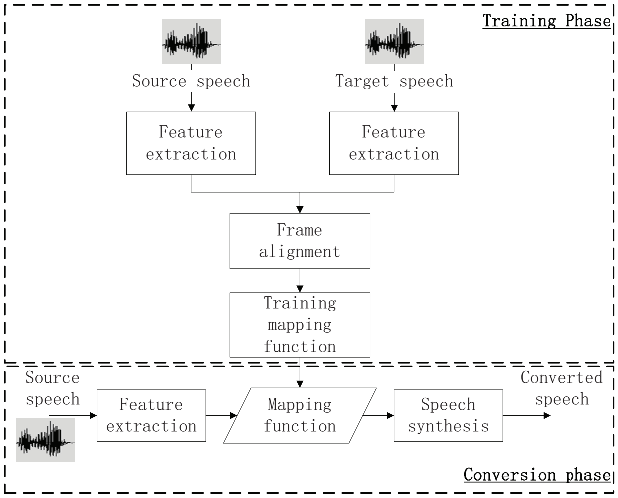

Figure 1: Diagram of the voice conversion system.

consists of 6760 gender-matched verification trials (3978 genuine and 2782 impostor) from 504 target speakers as shown in Table 1. We follow the same evaluation rules as in the NIST 2006 SRE specifications.

In the **spoofing corpus**, the speaker models are the same as in the baseline corpus but the test utterances are processed through voice conversion. The 3978 genuine trials are kept untouched but the 2782 impostor trials undergo voice conversion. Note that voice conversion operates on a pair of speakers (the source and the target). This implies that, unlike in the typical NIST SRE tasks where the *same* test utterances and speaker models are re-used in multiple trials, we need to train different conversion function for each speaker pair in the trial list. As the speech files in SRE 2006 have an average duration of 5 minutes (of which about half contains speech), this poses a computational challenge. This is the main reason why our task contains significantly less verification trials compared to recent NIST SRE tasks. Similar to previous studies [5, 20], the utterances used for training the speaker enrollment models and voice conversion functions are disjoint. We utilize data from the 3- and 8-conversion training sections of the SRE 2006 to train the conversion functions.

3. Voice Conversion Methods

3.1. Stochastic Conversion Function

The mainstream voice conversion method is based on Gaussian mixture models (GMMs) [13, 14, 12]. In this study, we use *joint density GMM* voice conversion method proposed originally in [21]. It is described as follows. Consider frame-aligned sequences of training vectors from the source ($\mathbf{x}$) and the target ($\mathbf{y}$) speakers:

$$\begin{aligned}\mathbf{x} &= [\mathbf{x}_1^\top, \mathbf{x}_2^\top, \dots, \mathbf{x}_t^\top, \dots, \mathbf{x}_T^\top]^\top \\ \mathbf{y} &= [\mathbf{y}_1^\top, \mathbf{y}_2^\top, \dots, \mathbf{y}_t^\top, \dots, \mathbf{y}_T^\top]^\top,\end{aligned}$$

where $\top$ denotes vector transpose. The vectors are stacked at the frame level into joint vectors $\mathbf{z}_t = [\mathbf{x}_t^\top, \mathbf{y}_t^\top]^\top$. The joint probability density of the source and target feature vectors is modeled by a GMM,

$$P(\mathbf{z}_t|\lambda^{(z)}) = \sum_{m=1}^{M} w_m^{(z)} \mathcal{N}(\mathbf{z}_t|\boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}),$$

where $\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}$ and $\boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}$ are the mean vector and covariance matrix of the multivariate Gaussian density $\mathcal{N}(\mathbf{z}_t|\boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$, respectively. The prior probabilities $w_m^{(z)}$ sum up to unity. The GMM parameters $\lambda^{(z)} = \{w_m^{(z)}, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)} | m = 1, 2, \dots, M\}$ are estimated in maximum likelihood sense using the well known expectation maximization (EM) algorithm. Here, we use $M = 8$ Gaussians with full covariance matrices. In the conversion phase, given a novel source speaker vector ($\mathbf{x}$), the trained joint density model is used for predicting the target speaker vector $\hat{\mathbf{y}}$ as,

$$\begin{aligned}\hat{\mathbf{y}} &= F(\mathbf{x}) = \mathsf{E}(\mathbf{y}|\mathbf{x}) \\ &= \sum_{m=1}^{M} p_m(\mathbf{x})(\boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)}(\boldsymbol{\Sigma}_m^{(xx)})^{-1}(\mathbf{x} - \boldsymbol{\mu}_m^{(x)})),\end{aligned}$$

where $p_m(\mathbf{x}) = \frac{w_m \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^{xx})}{\sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^{xx})}$ is the posterior probability of source vector $\mathbf{x}$ originating from the m^{th} Gaussian. We use the above procedure to convert spectral parameters. For the fundamental frequency (F0), conversion is done by equalizing the means and variances of the source and target log-F0 distributions as is commonly done.

3.2. Non-Parallel Frame Alignment Using VQ Mapping

We now discuss how to align the training vectors as required by the stochastic conversion framework. Typically one uses a set of *parallel* training utterances from the source and the target speakers. That is, same text passages read by both speakers. These training utterances would then be time-aligned using, for instance, dynamic time warping (DTW). The corpus used in this study, unfortunately, consists of conversational telephone speech without parallel utterances. Thus, we have to resort to nonparallel alignment methods [22, 23, 24].

In preliminary experiments, we implemented the nonparallel alignment method of [23] which simultaneously finds frame alignment and conversion function in multiple iterations. This led to good conversion quality but with high computational load. Hence, we ended up using faster vector quantization (VQ) based approach proposed in [22]. For completeness, we summarize the approach here.

1. Let $\mathcal{X} = \{\mathbf{x}_i\}$ and $\mathcal{Y} = \{\mathbf{y}_j\}$ be the alignment vectors of source and target, respectively. Using K-means, we train two codebooks $C^{(x)} = \{\mathbf{c}_1^{(x)}, \dots, \mathbf{c}_K^{(x)}\}$ and $C^{(y)} = \{\mathbf{c}_1^{(y)}, \dots, \mathbf{c}_K^{(y)}\}$ of K centroid vectors using $\mathcal{X}$ and $\mathcal{Y}$, respectively.
2. Create an index map $g(k)$ from the source to target clusters by the nearest neighbor rule, i.e.

$$g(k) = \arg\min_{1 \le r \le K} \|\mathbf{c}_k^{(x)} - \mathbf{c}_r^{(y)}\|^2, \quad k = 1, \dots, K.$$

Table 2: Performance of five different speaker recognition systems under voice conversion on the **spoofing** corpus.

Voice conversion	Equal error rates (EER %)					100 × MinDCF				
	GMM-UBM	VQ-UBM	GLDS-SVM	GMM-SVM	GMM-JFA	GMM-UBM	VQ-UBM	GLDS-SVM	GMM-SVM	GMM-JFA
None (Baseline)	7.63	7.56	7.16	3.74	3.24	3.54	3.07	3.03	1.70	1.57
MCEP, uni-frame align.	24.99	22.62	25.17	12.58	7.61	8.44	7.62	9.57	4.91	3.49
MCEP, tri-frame align.	24.49	20.74	23.41	12.88	7.40	8.29	7.11	9.49	4.75	3.41
LSP, uni-frame align.	21.90	19.81	18.70	10.81	6.48	7.58	6.25	6.86	4.17	2.72
LSP, tri-frame align.	21.07	19.16	17.15	10.81	6.30	7.31	6.01	6.65	3.88	2.65

3. For each source training vector $\mathbf{x}_i \in \mathcal{X}$, let k^* to be the index of it's nearest centroid in the source codebook. That is, $k^* = \arg\min_{1 \leq k \leq K} \|\mathbf{x}_i - \mathbf{c}_k^{(x)}\|^2$.
4. The paired target vector corresponding to $\mathbf{x}_i$ is the nearest neighbor of $\mathbf{x}_i$ in the vectors assigned to cluster $g(k^*)$ in the target codebook $C^{(y)}$.

For alignment, we use, in fact, different features from the actual conversion features (subsection 3.3). For the alignment, we use mel-frequency cepstral coefficients (MFCCs) due to their success in speech recognition. We extract 12 MFCCs with deltas (not including the energy coefficient). Energy-based voice activity detection (VAD) is performed since non-speech frames degrade conversion quality [25]. Utterance-level cepstral mean and variance normalization (CMVN) are used for speaker and channel normalization. As alternative alignment features, we also consider *tri-frame* alignment method [24] which expands the left and right acoustic contexts and was shown to work well on the non-telephony CMU Arctic data.

3.3. Conversion Features

The sampling rate of our speech files is 8 kHz. The speech signal is windowed in 25 ms window with a 5 ms shift and only the detected speech frames undergo conversion. We consider two spectral parameterizations, 30 mel-cepstrum coefficients (**MCEP**) and line spectrum pairs (**LSP**). The features are extracted using the SPTK tool [26]. F0 values are automatically extracted using the RAPT algorithm [27]. After conversion, SPTK tool is also used to synthesize speech.

4. Speaker Verification Systems

In the experiments, we consider five speaker verification systems of varying complexity. All systems use the same acoustic front-end consisting of 12 MFCCs with Δ and Δ^2 coefficients computed via 27-channel mel-frequency filterbank. RASTA filtering, voice activity detection (VAD) and utterance CMVN are applied as postprocessing. The energy VAD decisions of test segments are derived from the original baseline corpus. In the evaluation, we consider equal error rate (EER) and MinDCF (using the cost parameters in the SRE 2006 plan).

GMM-UBM: This is the standard Gaussian mixture model with universal background model (UBM) [15]. We train the UBM with 2048 Gaussians using EM algorithm from the NIST 2004 SRE corpus. We adapt the target speaker models using maximum *a posteriori* (MAP) adaptation of the UBM means.

VQ-UBM: Similar to GMM-UBM, we model each speaker using a vector quantizer codebook of 2048 code vectors trained using MAP adaptation [16]. The background utterances are the same as for GMM-UBM.

GLDS-SVM: Generalized linear discriminant sequence (GLDS) kernel support vector machine [17] uses 3^{rd} order monomial expansions, leading to 9139-dimensional polynomial supervectors per utterance. Speaker models are trained using LibSVM [28]. The same background utterances are used as for the previous two systems.

GMM-SVM: In the GMM supervector method [18], we first train a UBM with 512 Gaussians. We then adapt utterance GMM mean supervectors of dimensionality $36 \times 512 = 18432$. These are compensated with NAP [19] and used for target speaker model training using LibSVM. The match scores are additionally normalized using ZT-norm. NIST SRE 2004, SRE 2005 and MIXER 5 data are used for training UBM, NAP, cohort models and in SVM background.

GMM-JFA: GMM-JFA builds up on joint factor analysis (JFA) [2] for intersession and speaker variability compensation. Similar to GMM-SVM, it uses 512 Gaussians but TZ-norm for score normalization. Same datasets as for the previous system, plus additionally Switchboard corpus, are used in training.

5. Results

The results are given in Table 2. Considering baseline accuracy, GMM-SVM and GMM-JFA recognizers outperform the other three lightweight recognizers as expected. When voice conversion is introduced, all recognizers are seriously damaged. The relative increase in EER for GMM-UBM, VQ-UBM and GLDS-SVM are 3-fold or more for the uni-frame MCEP conversion. Even the EER of GMM-JFA is more than doubled. Regarding MinDCF, GLDS-SVM experiences the worst degradation (more than 3-fold increase). For the other systems – including GMM-JFA – MinDCF values are more than doubled.

From the conversion methodology point of view, the mel-cepstrum based method systematically outperforms the LSP conversion since it gives higher speaker verification error rates. This might be because the recognizers also use MFCCs; the simulated voice conversion intruder here has knowledge on the recognition system [9]. The uni-frame alignment method, in turn, systematically outperforms tri-frame alignment. This is different from our earlier result [24] on wideband microphone data (CMU Arctic) using small number of speakers. The current study utilizes larger 8 kHz telephony data containing significant channel effects and some additive noise which may explain the difference.

Increase in EER and MinDCF reflect loss in discriminatory information but might still give too optimistic viewpoint when the decision thresholds are trained on original data and applied to converted data. Therefore, we choose the two best-performing recognizers, GMM-SVM and GMM-JFA, and set the decision threshold to the EER threshold on the baseline corpus. We then measure the false acceptance rate on the spoofing corpus using the most successful MCEP conversion method with uni-frame alignment. According to Table 3, the false acceptance rates increase by factors of approximately 11:1 and 5:1 for the GMM-SVM and GMM-JFA systems, respectively. Even

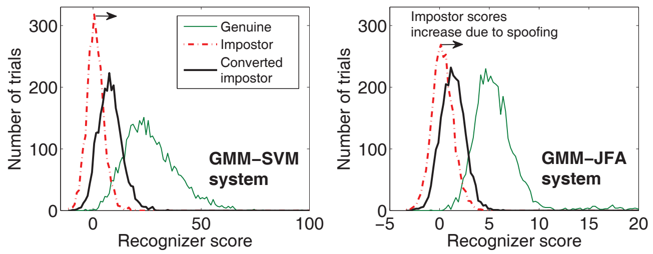

Figure 2: Score distributions before and after spoofing attack.

though the baseline FARs are close to each other, the GMM-JFA system experiences less degradation. The recognition system score distributions in Fig. 5 also indicate that GMM-JFA impostor score distribution is less affected by spoofing.

Table 3: Effect of spoofing to false acceptance rates (FAR, %). Decision threshold is set to EER point on the baseline corpus.

	GMM-SVM	GMM-JFA
Baseline	3.74	3.24
Spoofing (MCEP, uni-frame)	41.54	17.33

6. Conclusions

We studied vulnerability of speaker verification systems against spoofing and disguise attacks. Our experiments indicate that a simple voice conversion system – even when trained using non-parallel alignment and telephone speech – was able to break down all the five recognizers considered. Thus, we **confirm** that the earlier findings on clean data hold also for telephone data. Importantly, our findings suggest that, even though GMM-JFA is mainly designed to handle intersession variabilities, it also shows higher resistance against spoofing in comparison to the simpler methods. We hypothesize that the voice conversion function introduces a form of channel shift to the features which is partly compensated for by the channel variability subspace model (this should be confirmed). Nevertheless, even the accuracy of JFA decreased to unacceptable level. We argue that a human listener would easily judge our converted samples to sound nonnatural and therefore, solutions for natural/non-natural speech discrimination are needed.

7. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[2] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," technical report CRIM-06/08-14, Montreal, CRIM, 2006.

[3] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification - a study of technical impostor techniques," in *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, Budapest, Hungary, September 1999, pp. 1211–1214.

[4] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA 10 workshop*, 2010, pp. 131–134.

[5] B. Pellom and J. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," Phoenix, Arizona, USA, March 1999, pp. 837–840.

[6] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using a HMM-based speech synthesis system," in *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Aalborg, Denmark, September 2001, pp. 759–762.

[7] P. DeLeon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 151–158 (paper 28).

[8] Q. Jin, A. Toth, A. W. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?" Proc. ICASSP 2008, March 2008, pp. 4845–4848.

[9] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech 2007 (ICSLP)*, Antwerp, Belgium, August 2007, pp. 2053–2056.

[10] Y. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the YOHO speaker verification corpus," in *Knowledge-Based Intelligent Information and Engineering Systems (KES 2005)*, Melbourne, Australia, September 2005, pp. 15–21.

[11] M. Farrús, M. Wagner, J. Anguita, and J. Hernando, "How vulnerable are prosodic features to professional imitators?" in *The Speaker and Language Recognition Workshop (Odyssey 2008)*, Stellenbosch, South Africa, January 2008.

[12] Y. Stylianou, "Voice transformation: A survey," in *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)*, Taipei, Taiwan, April 2009, pp. 3585–3588.

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

[14] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Systems*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[15] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[16] V. Hautamäki, T. Kinnunen, I. Kärkkäinen, M. Tuononen, J. Saastamoinen, and P. Fränti, "Maximum *a Posteriori* estimation of the centroid model for speaker verification," *IEEE Signal Processing Letters*, vol. 15, pp. 162–165, 2008.

[17] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, April 2006.

[18] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[19] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, USA, March 2005, pp. 629–632.

[20] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2000)*, vol. 2, Beijing, China, October 2000, pp. 302–305.

[21] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 285–288.

[22] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A first step towards text-independent voice conversion," in *Proc. Interspeech 2004*, Jeju, South Korea, October 2004, pp. 1173–1176.

[23] D. Erro and A. Moreno, "Frame alignment method for cross-lingual voice conversion," in *Proc. Interspeech 2007 (ICSLP)*, Antwerp, Belgium, August 2007, pp. 1969–1972.

[24] Z.-Z. Wu, T. Kinnunen, E. Chng, and H. Li, "Text-independent F0 transformation with non-parallel data for voice conversion," in *Proc. Interspeech 2010*, Makuhari, Japan, September 2010, pp. 1732–1735.

[25] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[26] "Speech Signal Processing Toolkit (SPTK) version 3.4," *Software available at http://sp-tk.sourceforge.net/.*

[27] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.