EXPLOITING MAGNITUDE AND PHASE SPECTRAL INFORMATION FOR CONVERTED SPEECH DETECTION

Maria Joana Correia [1], *Alberto Abad* [1,2], *Isabel Trancoso* [1,2]

[1] INESC-ID/Spoken Language Systems Laboratory, Lisbon, Portugal
[2] Instituto Superior Técnico, University of Lisbon, Portugal

ABSTRACT

Speaker verification systems have been shown to be vulnerable in situations where voice conversion techniques are used to try to fool them, evidencing an important security breach in these applications.

This work focuses on the development of a new converted speech detector able to robustly address this problem. The proposed detector uses four spectral features extracted from the magnitude and the phase spectrum of the speech signal. To evaluate the performance of the detector we use a subset of the core task of the NIST SRE2006 corpus as the natural data. The converted data was produced with two different voice conversion methods: Gaussian mixture model and unit selection, from other NIST SRE2006 conditions. The converted speech detector achieved a detection accuracy of 99.1% and 98.5% for natural and converted utterances, respectively.

Index Terms — Voice conversion, Speaker verification, Converted speech detection, Spoofing

1. INTRODUCTION

Voice conversion (VC) is any technique that aims at modifying the voice of a speaker, the *source* speaker, so that it sounds like it belongs to another speaker, the *target* speaker, without changing the linguistic content of the converted utterance. Presently these techniques are capable of generating reasonably natural sounding speech using small amounts of data from the target speaker. On the other hand, speaker verification (SV) consists of automatically deciding either to accept or to reject a claimed identity, based on a user's utterance [1]. It is possible to perform a spoofing attack against a SV system by using converted speech to try to fool it. Hence, systems using this type of technology can pose a risk to SV systems.

There are several studies confirming the vulnerability of SV systems against converted speech spoofing attacks [2][3][4]. All of these clearly show that the false acceptance rate (FAR) of the SV system increases significantly.

The most successful approach to manage the high FAR in spoofing situations is to include a converted speech detector as a post processing module for the SV system. Several authors have proposed converted speech detectors based on different features, such as prosodic [5], phase [6] and pitch pattern [7], among others.

In this paper we propose a new converted speech detector that uses four spectral features extracted from the magnitude and phase spectrum, allowing a more thorough characterization of the speech signal. We also adopt a modeling paradigm based on support vector machines (SVMs), as was introduced in our previous work [8].

This paper is organized as follows: Section 2 briefly describes the VC methods that were used to create the converted speech corpora in this work. In Section 3, we introduce the converted speech detector proposed in this work, discussing the modeling technique and feature parameterization it will use. The experiments, results and their discussion are presented in Section 4. Finally, we draw some conclusions in Section 5.

2. VOICE CONVERSION METHODS

In this study we consider converted speech achieved through one of two VC methods: GMM- or unit selection (US)-based. These are briefly described in Sections 2.1 and 2.2, respectively.

2.1 GMM-based voice conversion

One of the most popular methods for voice conversion was originally proposed by [9] and is based on the joint density Gaussian mixture model.

This model often requires N-dimensional time aligned acoustic features, $\mathbf{X} = [\mathbf{x}_1', \mathbf{x}_2', \dots, \mathbf{x}_N']'$, from the source speaker and $\mathbf{Y} = [\mathbf{y}_1', \mathbf{y}_2', \dots, \mathbf{y}_N']'$, from the target speaker, determined, for instance, by dynamic time warping (DTW). These can be combined in feature vector pairs $\boldsymbol{Z} = [\boldsymbol{z_1}', \boldsymbol{z}_2', \dots, \boldsymbol{z}_T']'$ where $\boldsymbol{z}_t' = [\boldsymbol{x}_n', \boldsymbol{y_m}']' \in \mathbb{R}^{2d}$.

In the GMM algorithm, the joint probability function of the acoustic features is defined as:

$$\boldsymbol{P(X,Y) = P(Z)} = \sum_{i=1}^{M} \boldsymbol{\alpha_i^{(z)} \mathcal{N}(z|\mu_i^{(z)};\Sigma_i^{(z)})}, \sum_{i=1}^{M} \boldsymbol{\alpha_i^{(z)} = 1, \alpha_i^{(z)} > 0}, \quad (1)$$

978-1-4799-7129-9/14/$31.00 ©2014 IEEE

where $\boldsymbol{\mu}_i^{(z)}$ is the mean and $\boldsymbol{\Sigma}_i^{(z)}$is the covariance of the *M*-variate normal distributions.

Parameters $\lambda^{(z)} = \left\{\alpha_i^{(z)}, \boldsymbol{\mu}_i^{(z)}, \boldsymbol{\Sigma}_i^{(z)} | i = 1,2,\ldots,M\right\}$ are estimated using the expectation maximization (EM) algorithm [10].

The mapping function [11] used to convert features from the source speaker to target speaker is given by:

$$F(\boldsymbol{x}) = E(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{M} p_i(\boldsymbol{x}) \left(\boldsymbol{\mu}_i^{(z)} + \boldsymbol{\Sigma}_i^{(yx)} \left(\boldsymbol{\Sigma}_i^{(xy)}\right)^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}_i^{(z)}\right)\right), \qquad (2)$$

where $p_i(\boldsymbol{x}) = \frac{\alpha_i \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i^x;\boldsymbol{\Sigma}_i^{xx})}{\sum_{k=1}^{L} \alpha_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k^x;\boldsymbol{\Sigma}_k^{xx})}$ is the posterior probability of the source vector belonging to the i^{th}mixture component.

2.2 Unit selection (US)-based voice conversion

Unlike GMM-based voice conversion that may require parallel data, US-based voice conversion does not require it; instead, it uses the target speaker's voice to directly synthesize new speech [12].

The goal of US is to, given a sequence of source speech features, $\boldsymbol{x}_1^M$, find the best fitting sequence of target speech features, $\boldsymbol{y}_1^M$, that minimizes the target cost (an estimate of the difference between the database unit u_m and the target t_m which it is supposed to represent), and the concatenation cost (an estimate of the quality of a join between the consecutive units u_{m-1} and u_m) [12]. The cost functions can be defined by interpreting the feature vectors as database units i.e. $t \coloneqq x$ and $u \coloneqq y$. The target vector sequence is given by:

$$y_1^M = \arg\min_{y_1^M} \sum_{m=1}^{M} \{\alpha S(\boldsymbol{y}_m - \boldsymbol{x}_m) + (1-\alpha) S(\boldsymbol{y}_{m-1} - \boldsymbol{y}_m)\}, \qquad (3)$$

where α is a parameter to adjust the tradeoff between fitting the accuracy of source and target sequences and the spectral continuity criteria.

3. CONVERTED SPEECH DETECTORS

A converted speech detector is a system which aims at discriminating natural and converted speech. It was introduced only a few years ago as a possible solution to address the security issues posed by converted speech to SV systems.

3.1 Modeling technique

The converted speech detectors previously presented in the literature are based on the GMMs modeling paradigm. However, natural and converted discrimination is a binary task; as such, we considered studying the performance of a discriminative approach to address it in our previous work [8]. The preliminary experiments carried out in that work showed that a converted speech detector based on SVMs yielded a better performance than the GMM-based ones. In this paper we expand the experiments carried out in our previous work and use exclusively SVMs as the modeling technique for the converted speech detectors. We briefly explain the concept of SVM in the Section 3.1.1.

3.1.1.SVM

Given a training set of labeled, two-class examples, an SVM estimates a hyperplane that maximizes the separation of the two classes, after transforming it to a high dimensional space via Kernel function. SVMs are constructed as a weighted sum of a kernel function:

$$f(x) = \sum_{i=1}^{L} \alpha_i t_i K(x, v_i) + k, \qquad (4)$$

where x is the input data, L is the number of support vectors, α_i and k are training parameters, v_i are the support vectors, obtained via an optimization process.

Traditionally, the output of an SVM for inputted test data is a predicted label, i.e. 0 or 1. This label is assigned depending on which side of the separating hyperplane the test features fall on. However, in this study we opted to use the distance to the hyperplane as the SVM output. The reasoning behind this is to allow a finer score fusion. The distance to the hyperplane works as a natural way to weight the confidence of the predicted label.

3.2 Feature parameterization

Most of speech related applications, including SV and VC systems, use features to characterize the speech signal extracted at frame level from the magnitude spectrum. It is known that the short-term magnitude spectrum carries most of the information relative to the acoustic cues of speech, which is enough information to characterize the speech and the speaker in most situations. However, information relative to high levels cues is mostly lost. In this study we intent to perform a more thorough characterization of the speech signal, so we use four different spectral features that capture information from the magnitude and the phase spectrum and from the short- and the long-term.

3.2.1. Short-term magnitude spectrum information

In this study we adopted the MFCCs, a facto standard in speech applications, as the feature to capture short-term magnitude spectrum information.

The MFCCs feature extraction process, for a given speech frame, $x(n)$, can be summarized as follows:

1. Computing the fast Fourier transform (FFT) $X(\omega)$, of $x(n)$.
2. Computing the power spectrum $|X(\omega)|^2$.
3. Applying a Mel-frequency filter bank to the power spectrum $|X(\omega)|^2$ to obtain the filter-bank energies.
4. Applying discrete cosine transform (DCT) to the log-scale filterbank energies to compute the MFCCs.

3.2.2. Short-term phase spectrum information

In order to extract features derived directly from the phase spectrum of a speech signal, it is necessary to compute the unwrapped phase [14]. An alternative that is computationally simpler is using the group delay function phase spectrum (GDFPS) [15], which has the additional advantage of reducing the effects of noise.

The GDFPS is a measure of non-linearity of the phase spectrum [16] and is defined as the negative derivative of the phase spectrum with respect to the frequency:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2}, \tag{5}$$

where $\mathrm{X}(\omega)$ and $\mathrm{Y}(\omega)$ are the FFT of $x(n)$ and $nx(n)$, $\mathrm{X_R}(\omega)$, $\mathrm{X_I}(\omega)$, $\mathrm{Y_R}(\omega)$ and $\mathrm{Y_I}(\omega)$ are the real and imaginary part of $\mathrm{X}(\omega)$ and $\mathrm{Y}(\omega)$, respectively.

Given a speech signal, the computation of group delay cepstral coefficients (GDCC) for each speech segment, $x(n)$, of length t, updated every $\frac{t}{2}$ was achieved as follows:

1. Computing the FFT $X(\omega)$ and $Y(\omega)$ of $x(n)$ and $nx(n)$, respectively.
2. Computing the GDFPS as in Equation (5).
3. Applying a c filter Mel-frequency filter-bank to the MGDFPS to obtain c filter-bank coefficients.
4. Applying the DCT to the c filter-bank coefficients to obtain p GDCC.

The resulting p GDCC are used as feature vectors.

3.2.3. Long-term magnitude spectrum information

To capture the correlation between frames and the temporal characteristics of features trajectories in the magnitude spectrum, one can compute the magnitude modulation (MM) features [13] as follows:

1. Dividing the power spectrogram into n frame segments with a $m < n$ frame overlap.
2. Applying a c filter Mel-filterbank to the spectrogram to obtain the filter-bank coefficients, forming a $c \times n$ matrix.
3. Applying mean variance normalization (MVN) to the trajectory of each filter-bank.
4. Computing the p-point FFT of the c normalized trajectories.
5. Concatenating every modulation spectra to form a $\frac{cp}{2}$ coefficients modulation supervector (the second half of point of the FFT is ignored given its symmetric nature).
6. Applying principal component analysis (PCA) to the modulation supervector to reduce dimensionality and eliminate dimensions with high correlation. Kept the k projected dimensions with the largest associated variance.

The MM feature vectors are then k-dimensional and are used as feature vectors.

3.2.4. Long-term phase spectrum information

Information relative to the long term phase spectrum can be extracted in a similar fashion than the long-term magnitude spectrum information. By following the process described in Section 3.2.3, but replacing the magnitude spectrogram by a group delay function phase spectrogram one can extract the phase modulation (PM) features [13], which can be used as feature vectors.

3.2.5 Compact feature representation

The usual feature representation for an utterance with N voiced frames is a matrix of $N \times C$ coefficients, where C is the number of coefficients of the feature vectors characterizing each frame.

Alternatively to this consuming, full representation of information, we proposed in [8] the use of a lighter alternative. To use this light feature representation, given an utterance, one should perform the selected feature extraction process as normally and obtain the $N \times C$ matrix. Then, compute the mean and standard deviation of each of the C coefficients over the whole utterance and form a new vector of length $2C$. This vector will compactly represent the utterance and should be used as the new feature vector.

Comparatively to the full representation we decrease the number of feature vectors approximately 10^4 times per minute of speech, assuming the features are extracted every 10ms.

The VC process usually introduces systematic artifacts in the converted utterance, which are reflected in the features characterizing. The converted features will be shifted to a range of values not typical of the natural features.

An advantage of this light representation is that it reduces the training time of the model from several hours to a few seconds.

3.2 System combination

The features extracted from short-time spectral frames carry complementary information to the long term-features [6], as well as features extracted from the magnitude and the phase spectrum. In previous work [8] we performed some preliminary experiments showing the usefulness of fusing the scores of two detectors using complementary features. Supported by the improved performance of the fused detectors, we further those experiments in this work by fusing the scores of four detectors, each using MFCCs, GDCCs MM or PM features.

The fusion function of the sub-systems is based on linear logistic regression (LLR), where the fused scores, l, are obtained as follows:

$$l = \sum_i \alpha_i s_i + b, \tag{6}$$

where α_i is the weight for the sub-system i and b is the offset. The parameters were trained in a development data set using a sort of 2-fold cross-validation [17]: development data is randomly split in two halves, one for parameter estimation and the other for assessment. This process can be repeated using n different random partitions and the mean of the systems' fusion parameters can be computed.

4. EXPERIMENTS AND RESULTS

4.1 Corpora

In this study we used four main speech corpora, two of them comprised of natural speech data, one of GMM-based converted speech data and one of US-based converted speech data.

For the natural data we used a subset of the training data of the NIST SRE 2006 1conv4w condition. It consisted of 300 speech files with five minutes, including silence. Of those, 150 were of a female speaker and 150 of a male speaker. The length of the files roughly halved after silence removal via voice activity detection (VAD) algorithm. We also used test data from the NIST SRE2006 1conv4w condition, randomly selecting 3647 files, of 504 unique speaker, of which 298were females and 206 males.

The converted speech used data from the NIST SRE2006 3conv4w and 8conv4w training Sections as source data and the conversion matched randomly chosen, same gender speakers from the 1conv4w of the NIST SRE2006. Hence the converted utterances also 5 minutes long, including silence. Two VC methods (GMM-based and US based) were used to perform the conversion, and resulted on 2747 and 2748 files, respectively.

Table 1 summarizes the number of files and the types of available corpora.

Natural speech		**Converted seech**	
SRE2006 1conv4w train	*SRE2006 1conv4w test*	*GMM-based*	*US-based*
300	3647	2747	2748

Table 1. Corpora used in the experiments

4.2 Converted speech detection with one feature set

Here we investigate the performance of converted speech detectors using SVM as the modeling technique, a compact feature representation and considering only one of the features described in Sections 3.2.1 to 3.2.4 per detector. In total we trained four independent converted speech detectors.

Regarding feature dimensionality, the MFCCs were 12 dimensional features vectors, excluding the 0th coefficient and without temporal derivatives. The GDCCs were also were 12 dimensional features vectors, excluding the 0th coefficient and without temporal derivatives. The MMs and the PMs were 10 dimensional feature vectors. The 10 projected dimensions account for more than 97.0% of the total variance in both cases.

As the train data for the natural speech model we use the train data available from the 1conv4w NIST SRE2006 corpus, totaling 300 speech files. The converted speech model was trained with 150 randomly chosen files of each of the converted speech corpora. We trained a single classifier with examples of GMM and US-based converted speech in order to provide broader examples of non-natural speech and make the converted speech model more robust.

The four converted speech detectors were tested against the remaining data not used for training: 3647 files from the test 1conv4w condition of the NIST SRE2006 corpus, 2597 GMM converted speech files and 2589 US based converted speech files.

The chosen performance metric for these experiments was the detection accuracy rate. It was computed by establishing the decision threshold at zero. Given that the output score of the detectors was the distance to the hyperplane instead of the predicted label, it was also possible to plot a detection-error trade-off (DET) curve of the performance of the detectors. Figure 3 shows the DET curve for the performance of the four independent detectors. The miss rate corresponds to the rate of converted trials that are misclassified as being natural and the false acceptance rate corresponds to the rate of natural trails that are falsely accepted as converted. Table 2 summarizes the performance of the detectors in detection accuracy for natural and converted trials when the decision threshold is set at zero.

The scores distributions, shown in Figure 1, illustrate how the artifacts present in the converted speech shift the scores of converted trials to a typically non-natural range of values. This is particularly evident for the detector using the GDCCs.

The best performing detector was clearly the one using GDCCs achieving an accuracy rate of 97.9% and 98.1% for natural and converted trials, respectively. The remaining detectors yielded poorer performances.

We note that the presented performance of the detector when faced against converted trials is a pooling of the performances of the GMM-based and US-based converted trials. We chose to pool the results for the sake of simplicity and because of the similar performances we obtained with the two conversion methods.

Feature	**Acc. % for trials**	
	Natural	*Converted*
MFCC	72.9	69.6
GDCC	97.7	97.9
MM	79.7	80.8
PM	80.0	80.8

Table 2. Performance, in accuracy rate, of the converted speech detectors against natural and converted data

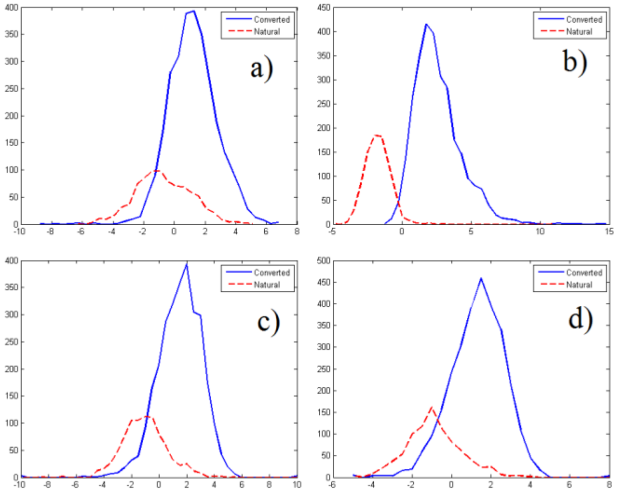

Figure 1. Score distributions for the converted speech detectors using a) MFCCs, b) GDCCs, c) MM features and d) PM features

4.3 Converted speech detection with multiple features

With the goal of making the converted speech detection task more robust, and considering the complementary information existent in the features extracted from the magnitude and phase spectrum, here we focused on performing the fusion of the four detectors described in Section 4.2.

The fusion was performed at score level, using the fusion algorithms implemented in the Bosaris toolkit for Matlab [18], which performs LLR to fuse multiple sub-systems of binary classification. The fusion parameters were trained with 10 iterations of the 2-fold cross-validation process.

The performance of the fused detector in accuracy rate is presented in Table 3. The decision threshold was kept at zero, as in the previous experiments.

Figure 2 shows the score distribution of the natural and the converted trials. From it, we can observe that the separation of the natural and converted scores distributions is more pronounces than in any of the previous detectors.

Figure 3 also shows the DET curve of the performance of the fused converted speech detector.

The fusion of the four sub-detectors allowed an improvement of the performance of the new detector. Comparatively to the detector using exclusively the GDCCs, the accuracy detection rate for natural trials showed an absolute improvement of 1.4%, achieving 99.1% accuracy; the absolute improvement of the accuracy detection rate for converted trials was of 0.6%, yielding 98.5% accuracy.

Acc. % for trials	
Natural	*Converted*
99.1	98.5

Table 3. Performance, in accuracy rate, of the fused converted speech detector against natural and converted data

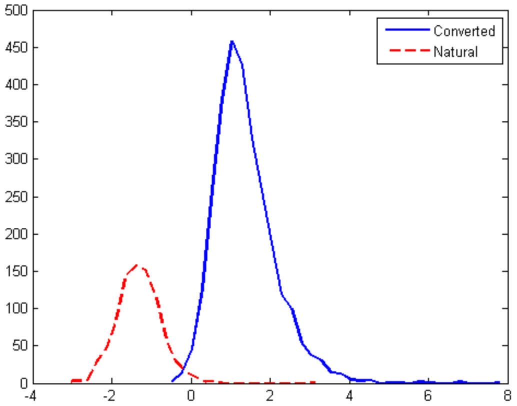

Figure 2. Score distributions of the natural and converted trials after the fusion of the four sub-detectors

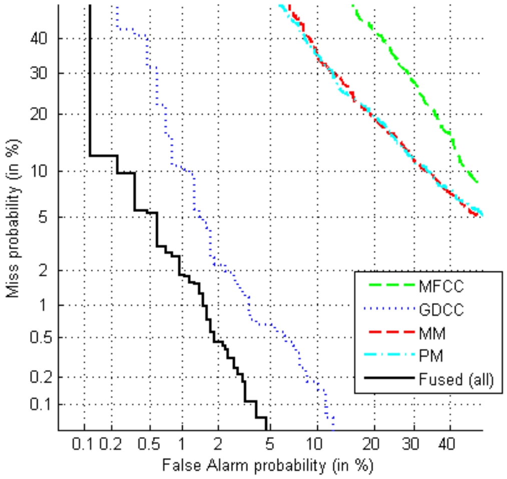

Figure 3. DET curve of the performance of the four simple converted speech detectors and the converted speech detector with the fused scores

5. CONCLUSION

In this paper we continued our previous work on addressing the security issue of speaker verifications systems when facing converted speech spoofing attacks. We proposed a new converted speech detector, based on the fusion of four sub-systems using features derived from the magnitude and the phase spectrum, both short- and long-termed. With this we were able to achieve a more thorough characterization of the speech signal, making the detection of the artifacts characteristic of converted speech easier, hence the discrimination between natural and converted speech more robust. We tested our converted speech detector against natural and converted speech data and achieved a detection accuracy of 99.1% for natural speech trials and 98.9% for converted speech trials. The performance of this detector using four sub-systems was better than the performance of any of the sub-detectors individually and was also an improvement from our previous work.

6. ACKNOWLEDGMENT

The authors would like to thank Zhizheng Wu and Haizhou Li of the Nanyang Technological University, Singapore for providing the GMM-based and US-based converted speech corpora, and for their helpful suggestions. This work was supported by FCT grants PTDC/EIA-CCO/122542/2010 and PEst-OE/EEI/ LA0021/2013, and EU Cost Action1206 De-identification for privacy protection in multimedia content.

7. REFERENCES

[1] J.P. Campbell Jr, “Speaker recognition: A tutorial,” *Proceedings of the IEEE,* vol. 85, no. 9, pp, 1437-1462, 1997

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

[3] Q. Jin, A. Toth, A. W. Black, and T. Schultz, “Is voice transformation a threat to speaker identification?,” *Proceedings of ICASSP 2008*, pp. 4845–4848, 2008.

[4] Q. Jin, A.R. Toth, T. Schultz, and A.W. Black, “Voice convesion: Speaker de-identification by voice transformation,” *Proceedings of ICASSP 2009,* pp. 3909–3912, 2009.

[5] A. Ogihara and A. Shiozaki, “Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences,* vol. 88, no. 1, pp. 280–286, 2005.

[6] Z. Wu, T. Kinnunen, E. Chng, H. Li, and E. Ambikairajah, “A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case,” APSIPA ASC 2012, 2012.

[7] Z. Wu, E. Chng, and H. Li, “Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition”, Interspeech 2012.

[8] M. J. Correia, A. Abad and I. Trancoso, “Preventing converted speech spoofing attacks in speaker verification,” *Proceedings of MIPRO 37th International Conference,* Opatija, Croatia, 2014.

[9] Kain, and M.W. Macon, “Spectral voice conversion for text-to-speech synthesis,” *Proceedings of ICASSP 1998*, vol. 1, pp. 285–288, 1998.

[10] D. Reynolds, “Automatic speaker recognition using gaussian mixture speaker models,” *The Lincoln Laboratory Journal,* vol. 8, no. 2, pp. 173-192, 1995.

[11] Y. Stylianou, O. Cappé, and E. Moulines, “Statistical methods for voice quality transformation,” *EUROSPEECH*, 1995.

[12] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, “Text-Independent Voice Conversion Based on Unit Selection,” *Proceedings of ICASSP 2006*, vol.1, pp.I-I, 2006.

[13] Z. Wu, X. Xiao, E. Chng, and H. Li, “Synthetic speech detection using temporal modulation feature”, *Proceedings of ICASSP2013*, vol. 1, no. 1, pp.7234,7238, 2013.

[14] J. Tribolet, “A new phase unwrapping algorithm,” *Acoustics, Speech and Signal Processing, IEEE Transactions* on , vol.25, no.2, pp.170-177, 1977.

[15] B. Yegnanarayana, J. Sreekanth, and A. Rangarajan, “Waveform estimation using group delay processing,” *Acoustics, Speech and Signal Processing, IEEE Transactions* on , vol.33, no.4, pp. 832- 836, 1985.

[16] L. D. Alsteris, and K. Paliwal, “Short-time phase spectrum in speech processing: A review and some experimental results,” *Digit. Signal Process*, vol. 17, no. 3 pp. 578-616, 2007.

[17] R. Fuentes “The BLZ Submission to the NIST 2011 LRE: Data Collection, System Development and Performance,” *Proceedings of Interspeech*, 2012.

[18] “Bosaris toolkit [software package],” WWW page, July 2014, https://sites.google.com/site/bosaristoolki