Simultaneous Utilization of Spectral Magnitude and Phase Information to Extract Supervectors for Speaker Verification Anti-spoofing

Yi Liu[1], *Yao Tian*[1], *Liang He*[1], *Jia Liu*[1], *Michael T. Johnson*[2]

[1]Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Speech and Signal Processing Lab, Department of Electrical and Computer Engineering,
Marquette University, Milwaukee, WI 53233, USA

{liuyi12, tianyao11}@mails.tsinghua.edu.cn, {heliang, liuj}@tsinghua.edu.cn,
michael.johnson@marquette.edu

Abstract

Protection from spoofing attacks is an essential component of speaker verification systems. This paper proposes a novel approach to detect such attacks by utilizing supervectors derived from spectral magnitude and phase information. Three countermeasures are chosen to represent these important information. To combine different countermeasures, score fusion and an anti-spoofing supervector (ASSV) are used. Experiments conducted on ASVspoof 2015 show that the combination of magnitude and phase information obtains relative 90% improvement in terms of the equal error rate (EER) compared to the best subsystem in the development set. The two systems can also be fused to further improve the performance. In addition to accuracy improvements, the new supervector framework is extensible and allows for a more flexible interface to the back-end classifier design.

Index Terms: speaker verification, anti-spoofing supervector, spectral magnitude and phase information, ASVspoof 2015

1. Introduction

Text-independent automatic speaker verification (ASV) plays an important role in biometric authentication. By virtue of new modeling methods, such as Joint Factor Analysis (JFA) and i-vector representations, ASV systems have become less susceptible to noise or channel effects. However, even state-of-the-art ASV systems are still quite vulnerable to deliberate spoofing attacks. Classical spoofing methods, such as impersonation, replay, speech synthesis, voice conversion (including artificial signal generation), can significantly increase the false acceptance rate of ASV systems [1].

To address this situation, recently several countermeasures have been proposed to make ASV systems more robust from spoofing [2]. Most current anti-spoofing algorithms are built on assumptions related to specific spoofing approaches. For instance, impersonators are considered to exhibit larger acoustic parameter variation than original speakers [3] and different presentation of channel noise and reverberation can be a mechanism to recognize recordings [4]. Speech synthesis and voice conversion are two of the most easily accessible and effective spoofing approaches, which have received the most attention [5]. Dozens of countermeasures have been proposed, including detecting phase information [6], modeling the pattern of long-term features like F0 statistics [7], and observing the texture of the spectrogram [8, 9].

These anti-spoofing algorithms have been shown to work well under individual experimental configurations. However, these experiments are evaluated across multiple databases, and use different performance metrics, which makes direct comparison very difficult, if not meaningless. Moreover, most proposed countermeasures just focus on one specific type of attack. This is a significant limitation, because in practical applications a wide variety of attacks would always be expected. One way to address this is through some generalized countermeasures. For example, higher-level features with one-class classifier [9] and temporal modulation features [10] are both effective to detect different spoofing attacks.

In this paper, three countermeasures based on spectral magnitude and phase information are extracted as supervectors and can be classified by support vector machines (SVM). In addition to score fusion, a new anti-spoofing supervector (ASSV) approach is presented to combine these countermeasures to detect diverse spoofing, including voice conversion and speech synthesis. We evaluate our method using the ASVspoof 2015 challenge [11] so that the results are comparable across participants.

The remainder of this paper is organized as follows. The three types of countermeasures we use are briefly introduced in Sections 2. Section 3 presents the details of supervector extraction and our ASSV structure. Experimental work is described in Section 4. Finally, Section 5 concludes the paper.

2. Spoofing countermeasures based on magnitude and phase information

Spoofing is a mechanism to trick an ASV system by imitating a target speaker. Although the imitation is not acoustically exact, spoofing can target the features and models used for speaker identification. Since such synthesized, converted or artificially generated utterances unavoidably change important parameters, making the speech acoustically "unnatural", it should be possible to explicitly discriminate such spoofed speech. Both magnitude and phase information in the frequency domain can be effectively utilized together to identify such differences. In this work we use three features representative of these important information. They are local binary patterns, modified group delay feature and cosine normalized phase feature.

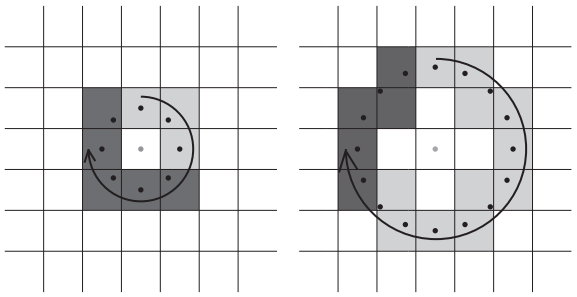
2.1. Local binary patterns

The spectro-temporal structure is an important property of an utterance. We hypothesize that many spoofing methods would alter the spectro-temporal structure, often referred to as "texture" [8]. To detect this disturbance, a well established approach called Local Binary Patterns (LBP) is used.

LBP was first proposed in texture recognition to represent local properties of a grey scale image. When dealing with a central pixel g_c in an image, the LBP operator characterizes P equally spaced pixels $(g_0, \ldots, g_{P-1})$ on a circle of radius R around g_c. The central pixel compares its gray-scale value with other neighbors and the sign of each comparison comprises the corresponding LBP index clockwise. The LBP operator can be expressed as:

$$\mathrm{LBP}_{P,R}(g_c) = \sum_{p=0}^{P-1} \mathrm{sign}(g_p - g_c) \cdot 2^p \tag{1}$$

where $\mathrm{sign}(x)$ denotes the sign function which equals 1 when $x >= 0$ and 0 otherwise. It is straightforward to see that there are 2^p possible raw LBP indices. Indices that can be equivalently represented by a circular bit-wise right shift are grouped together to remove the effect of rotation. Examples for $(P = 8, R = 1.0)$ and $(P = 16, R = 2.0)$ are shown as Figure 1.



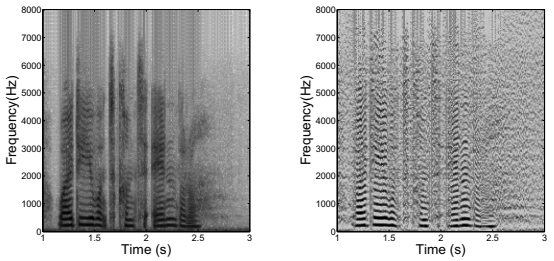
(a) $(P = 8, R = 1.0)$, $\mathrm{LBP} = (00011111)_\mathrm{b}$

(b) $(P = 16, R = 2.0)$, $\mathrm{LBP} = (0000000000011111)_\mathrm{b}$

Figure 1: *Local neighbors of the central pixels, and the corresponding binary LBP indices. The subscripts "b" denotes binary notation. Pixels with gray-scale values larger than the central pixels are marked with a deeper color.*

In [12], the authors discussed that the fundamental properties of image texture concentrate on certain patterns. These special patterns are defined as "uniform" LBPs, which have a common structure that only contains bit-wise transitions not larger than 2. Other nonuniform patterns are treated as a single miscellaneous group. After the original gray scale image is converted to LBP indices, the texture feature is extracted as the histogram of the uniform LBP over the whole sample.

We apply this concept to speech signal processing by using the spectrogram as acoustic representation, and treat it as a 2-D "image" so that the LBP operator can also be applied to estimate the feature texture. Hence, the texture of the spectral magnitude becomes a representative feature to discriminate spoofed speech. The same idea is stated in [8]. The feature they use is cepstral coefficients such as LPCC, while we find the less-processed FBank coefficients, which consist of mel-frequency filter-bank energies, give better performance in our experiments.



(a) Conventional power spectrogram

(b) Modified group delay spectrogram

Figure 2: *Comparison between power and modified group delay spectrograms over a two second interval of an utterance in ASVspoof 2015. The modified group delay spectrogram shows the same formant information as the power spectrogram, but also provides additional structure.*

2.2. Modified group delay feature

It has been demonstrated that the phase spectrum is useful for human auditory perception [13]. Most current spoofing algorithms do not retain the natural phase information, which makes them more vulnerable to detection in the phase domain as opposed to magnitude spectra.

Modified group delay features (MGDF) have been used in phoneme recognition for a long time [14]. Derived from the modified group delay function, MGDF provides meaningful phase information. Results have shown that MGDF is well suited to defend voice conversion attacks in contrast to conventional cepstral features [10].

Given a speech signal $x(n)$, the modified group delay function is expressed as:

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^2} \tag{2}$$

$$\tau_{\alpha,\gamma}(\omega) = \frac{\tau_p(\omega)}{|\tau_p(\omega)|} \cdot \left| \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}} \right|^\alpha \tag{3}$$

where $X(\omega)$ and $Y(\omega)$ are the short-time Fourier transform (STFT) of $x(n)$ and $nx(n)$ respectively, and the subscripts R and I denote the real and imaginary parts of a complex number. $|S(\omega)|^2$ is the smoothed version of $|X(\omega)|^2$, which can be obtained by simple cepstral smoothing or median filtering. Two parameters α and γ are introduced to further suppress the spiky nature of group delay spectrum. The difference between the power spectrum and the modified group delay spectrum is shown in Figure 2.

The discrete cosine transform (DCT) is applied to decorrelate the coefficients and obtain the final MGDF. The zero-th coefficient c_0 is ignored per [14].

2.3. Cosine normalized phase feature

Unlike MGDF, cosine normalized phase features (CNPF) use a more direct approach to extract phase information [15]. After the STFT is applied to a speech frame, the short-time phase spectrum $\psi(\omega)$ is first unwrapped to eliminate discontinuity. Cosine normalization is necessary to constrain the dynamic range to $[-1, 1]$. Then, a CNPF is extracted using a DCT applied to the cosine normalized phase spectrum.

3. Anti-spoofing supervector extraction

Motivated by the success of supervectors in speaker recognition [16], we use a supervector-based structure in our anti-spoofing system. Each utterance is converted to a single supervector, allowing flexibility for different classifiers.

To make the LBP features described in Section 2 more appropriate for spectrogram spoofing detection, we note some differences between traditional LBPs and those applied to spectrograms:

- Unlike images, spectral texture does not need rotation invariance, since rotation does not occur in speech-based spectrograms.
- The rows of feature texture are physically significant. If we use filter-bank energies as features, each row denotes the spectral character of a certain frequency range. Computing the histogram over the whole image would eliminate this useful information.

Taking these two points into account, we compute histograms across different rows of the unique LBPs without rotation invariance, as shown in Figure 3. Histograms are normalized individually to compensate for utterance duration and concatenated together to form the final normalized unique LBP (NULBP) supervector $\mathrm{SV}_{\mathrm{NULBP}}$.

In contrast to the NULBP feature, the MGDF and CNPF are extracted following a frame-by-frame fashion as conventional acoustic features. To extract supervectors for this kind of feature, a Gaussian mixture model (GMM) - universal background model (UBM) is first trained across all training speech. For each utterance, a GMM is adapted from the UBM using maximum a posteriori (MAP) adaptation:

$$g(\boldsymbol{x}) = \sum_{i=1}^{N} \lambda_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_i, \boldsymbol{\Sigma}_i) \tag{4}$$

where N denotes the mixture number, λ_i are the weights of the mixtures, $\mathcal{N}()$ indicates a Gaussian distribution, and $\boldsymbol{m}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and covariance of the i-th Gaussian. The means of the GMM are scaled [16]:

$$\hat{\boldsymbol{m}}_i = \sqrt{\lambda_i} \boldsymbol{\Sigma}_i^{-1/2} \boldsymbol{m}_i \tag{5}$$

and stacked to form the MGDF/CNPF supervector.

$$\mathrm{SV}_{\mathrm{MGDF/CNPF}} = \left[\hat{\boldsymbol{m}}_1^T, \hat{\boldsymbol{m}}_2^T, \ldots, \hat{\boldsymbol{m}}_N^T \right]^T \tag{6}$$

Although these three countermeasures are designed to detect a variety of spoofing attacks, they are dissimilar and there is not a simple way to merge them directly. If the parameters are carefully selected, system fusion among these three features (e. g. score-level fusion) would improve the performance; however, this approach sometimes requires more classifiers than available, and needs a development set as well. In this paper, we also concatenate these supervectors to yield what we term an anti-spoofing supervector (ASSV). The flowchart of ASSV extraction is demonstrated in Figure 3. Our ASSV is described as

$$\mathrm{ASSV} = \left[\mathrm{SV}_{\mathrm{NULBP}}^T, \mathrm{SV}_{\mathrm{MGDF}}^T, \mathrm{SV}_{\mathrm{CNPF}}^T \right]^T \tag{7}$$

This approach is extensible, so that an even larger ASSV can be created, if additional anti-spoofing features are available.

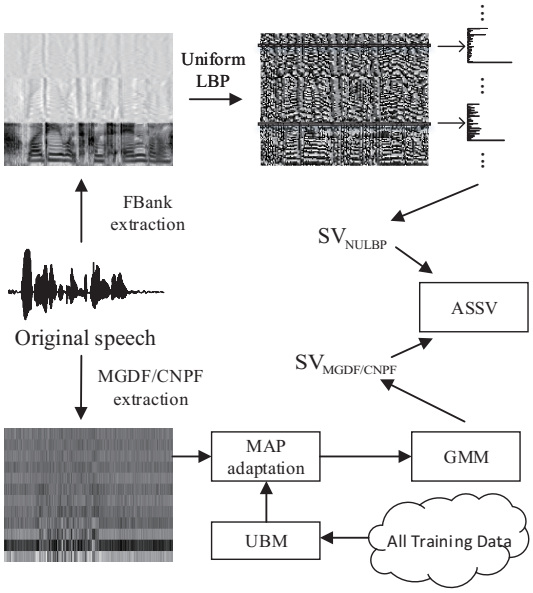


Figure 3: *The flowchart of supervector extraction and ASSV construction.*

4. Experimental work

4.1. ASVspoof 2015 challenge

The ASVspoof 2015 challenge was designed to evaluate the performance of state-of-the-art spoofing countermeasures. The database contains both genuine and spoofed speech. In the training and development set, three voice conversion and two speech synthesis algorithms are presented. These attacks are treated as *known attacks* and denoted as *S1-S5* in this challenge. To avoid the use of prior knowledge, five more algorithms were added to the evaluation set which were unseen during the development period, thus generalized countermeasures would be preferred. These additional spoofing algorithms contain *unknown attacks* which are denoted as *S6-S10*. More details about ASVspoof 2015 are described in [11].

4.2. Experiment setup

The NULBP is derived from 120-dimension FBank features: 40 mel-frequency filter-bank energies with their first and second derivatives. P and R are chosen to be 8 and 1.0, respectively. The miscellaneous group described in Section 2 is discarded, leaving 58 indices. Histograms for all coefficients except the first and last are computed, giving a total NULBP dimension of $58 * (120 - 2) = 6844$.

Both MGDF and CNPF use 12-dimension static features. We fix $\gamma = 1.2$ and $\alpha = 0.4$ in MGDF as described in [15]. Cepstral normalization is used on MGDF but not on CNPF in according to the actual experimental results. Each GMM consists of 512 mixture components. The UBMs are trained from all utterances in the training set. GMMs for each utterance are adapted using only the UBM means [16]. The means are stacked to form $512 * 12 = 6144$ length supervectors.

Each type of supervector is classified by a two-class SVM with linear kernel, and is considered as a separate *sub-system*. Our ASSV concatenates the above three supervectors thus the dimension would be $6844 + 6144 * 2 = 19132$. A single SVM

is able to discriminate the class of each ASSV. In contrast, the score fusion system is optimized in the development set. Although the standard way of score fusion is logistic regression [17], this paper uses grid search to tune the parameters for simplicity.

In comparison with our supervector method, the traditional GMM log-likelihood scoring is also implemented [10, 15]. In this approach, two GMMs are trained separately on genuine and spoofed speech. During the test phase, each utterance scores against these two GMMs to compute the likelihood ratio.

In our experiments, we strictly follow the common training condition of the ASVspoof 2015 challenge. Performance is reported for the development and evaluation sets. The results for the evaluation set were returned by the organizer of ASVspoof 2015. The equal error rate (EER) is used as the official primary metric for ASVspoof 2015.

4.3. Results

We first compare the performance of sub-systems in the development set using separate countermeasures. Both GMM log-likelihood scoring and supervector systems are applied to MGDF and CNPF. The results are shown in Table 1.

Table 1: *The performance of sub-systems in the development set.*

Sub-system	EER(%)
NULBP+SVM	0.858
MGDF+GMM	3.713
MGDF+SVM	2.602
CNPF+GMM	4.487
CNPF+SVM	4.403

Among the three countermeasures, NULBP achieves better performance than MGDF and CNPF. This suggests that the spectrogram texture effectively indicates spoofed speech. Meanwhile, phase information, i. e. MGDF and CNPF, is also useful. On the other hand, as is also typically the case for speaker verification, a supervector with linear kernel SVM consistently outperforms GMM log-likelihood scoring. It should be pointed out that, i-vector technique [18] was also explored in this challenge, but inferior performance was observed. So it is excluded and only the supervector structure is used in the remaining experimental comparisons.

In order to combine the ability of the three countermeasures, both score fusion and the proposed ASSV approach are evaluated. We also combine the two fusion systems to further boost performance. The results in the development and evaluation sets are shown in Table 2.

Table 2: *The performance of combined systems.*

Combined System	EER(%)			
	dev	eval known	eval unknown	eval avg
Score fusion (Grid Search)	0.058	0.104	6.775	3.439
ASSV	0.117	0.159	6.227	3.193
Score fusion+ASSV	**0.025**	**0.059**	**6.114**	**3.086**

From Table 2, we find that although the MGDF and CNPF systems do not perform as well as NULBP, their inclusions in the score fusion and ASSV systems reduce the EER from the best subsystem's 0.858% to 0.058% and 0.117%, leading to 93% and 86% relative improvements in the development set. This phenomenon shows the importance of the complementary information contained in spectral magnitude and phase.

For known attacks (*S1*-*S5*) in the development and evaluation sets, score fusion achieves better results than the proposed ASSV approach. Our hypothesis for this is that ASSV simply concatenates all supervectors together, without any weighting technologies, even though the discriminative abilities of different supervectors are unbalanced. The final system with both the score fusion and ASSV combined together gives additional performance improvement.

However, when it comes to unknown attacks (*S6*-*S10*), both systems perform considerably worse, and achieve basically the same EERs. Table 3 illustrates the performance of different attacks in the evaluation set. The results come from our original primary submission before the ASVspoof 2015 deadline.

Table 3: *Attack-dependent EER results in the evaluation set.*

Known	*S1*	*S2*	*S3*	*S4*	*S5*
EER(%)	0.173	0.610	0.319	0.289	0.399

Unknown	*S6*	*S7*	*S8*	*S9*	*S10*
EER(%)	0.906	0.242	0.417	0.246	28.581

It seems that *S6*-*S9* are relatively easy to detect because they are voice conversion algorithms using the same STRAIGHT vocoder with *S1*-*S4* [19]. The main error comes from *S10* which dwarfs all others by 2 orders of magnitude. *S10* denotes a text-to-speech technology that does not use a vocoder. The fact that no vocoder involved makes it sound more natural. The magnitude texture and phase information of such speech are unfamiliar to our classifiers as well, making it difficult to recognize. New approaches need to be developed to overcome this problem.

5. Conclusions

This paper presents a novel anti-spoofing system that combines several unique features based on both spectral magnitude and phase information. The supervector structure is effective at detecting attacks, and outperforms the conventional GMM log-likelihood scoring method. Experimental results show that, compared with the subsystems, score fusion and ASSV both greatly reduce the EER, by 93% and 86% in the development set, respectively. For known attacks, the fusion of these two systems lowers the EER further, while more countermeasures are needed to detect advanced unknown spoofing attacks. Overall the combination of spectral magnitude and phase features produces a significant improvement in spoofing detection, and the proposed ASSV framework is extensible and has the potential for a simple but flexible back-end design.

Future work includes incorporating a weighting algorithm to emphasize specific dimensions of ASSV in proportion to their discriminative ability, and enhancing ASSV with other powerful supervectors.

6. Acknowledgements

The authors want to thank Zhizheng Wu in University of Edinburgh for his timely reply to our ASVspoof 2015 submissions.

The work is supported by National Natural Science Foundation of China under Grant No. 61370034, No. 61403224 and No. 61273268.

7. References

[1] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, December 2012, pp. 1–5.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130 – 153, February 2015.

[3] T. B. Amin, J. S. German, and P. Marziliano, "Detecting voice disguise from speech variability: Analysis of three glottal and vocal tract measures," *Proceedings of Meetings on Acoustics*, vol. 20, no. 1, 2014.

[4] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, vol. 4, July 2011, pp. 1708–1713.

[5] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King, "SAS: A speaker verification spoofing database containing diverse attacks," in *INTERSPEECH*, 2015.

[6] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, 2012.

[7] P. L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *INTERSPEECH*, 2012.

[8] F. Alegre, R. Vipperla, A. Amehraye, and N. W. D. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *INTERSPEECH*, August 2013.

[9] F. Alegre, A. Amehraye, and N. W. D. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *6th IEEE International Conference on Biometrics: Theory, Applications and Systems(BTAS)*, September 2013.

[10] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7234–7238.

[11] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," 2015. [Online]. Available: http://www.spoofingchallenge.org/is2015_asvspoof.pdf

[12] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, Jul 2002.

[13] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153 – 170, 2005.

[14] H. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2003 IEEE International Conference on*, vol. 1, April 2003, pp. I–68–71.

[15] E. S. C. H. Zhizheng Wu, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *INTERSPEECH*, 2012.

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

[17] N. Brummer and D. van Leeuwen, "On calibration of language recognition scores," in *Speaker and Language Recognition Workshop, IEEE Odyssey*, June 2006, pp. 1–8.

[18] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[19] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187 – 207, April 1999.