Variants of Mel-frequency Cepstral Coefficients for Improved Whispered Speech Speaker Verification in Mismatched Conditions

Milton Sarria-Paja [1,2] and Tiago H. Falk [1]
[1]Institut National de la Recherche Scientifique (INRS-EMT), Montreal, Quebec, Canada
[2]Universidad Santiago de Cali, Cali, Colombia
milton.sarria00@usc.edu.co, falk@emt.inrs.ca

***Abstract*—In this paper, automatic speaker verification using normal and whispered speech is explored. Typically, for speaker verification systems, varying vocal effort inputs during the testing stage significantly degrades system performance. Solutions such as feature mapping or addition of multi-style data during training and enrollment stages have been proposed but do not show similar advantages for the involved speaking styles. Herein, we focus attention on the extraction of invariant speaker-dependent information from normal and whispered speech, thus allowing for improved multi vocal effort speaker verification. We base our search on previously reported perceptual and acoustic insights and propose variants of the mel-frequency cepstral coefficients (MFCC). We show the complementarity of the proposed features via three fusion schemes. Gains as high as 39% and 43% can be achieved for normal and whispered speech, respectively, relative to the existing systems based on conventional MFCC features.**

***Index Terms*—Whispered speech, speaker verification, fusion, i-vector extraction, MFCC.**

I. Introduction

Over the years, whispered speech has been the subject of different analyses ranging from perceptual to acoustic studies, in order to explore possible everyday applications. For example, perceptual studies have been conducted to characterize major acoustic differences between whispered and normal speech. Topics such as pitch perception and the correlation between perceived pitch and formant location have been studied, as well as the measurement of the formant shifts towards higher frequencies. Moreover, perceptual studies have suggested that whispered speech still conveys a significant amount of speaker identity and gender information [1], [2]. Acoustic studies, on the other hand, have corroborated and complemented perceptual findings. For instance, whispered speech has a lower and flatter power spectral density [3], and the duration of consonants in whispered speech are prolonged by about 10% relative to normally-voiced speech and the intensity of whispered consonants is lower by about 12 dB [4]. The above-mentioned insights have been used by the research community to tackle different challenges, such as reconstruction of normal speech from whispers [5], speech recognition [3], and speaker identification [6], [7], [8] with whispered speech.

This work was funded by the Natural Sciences and Engineering Research Council of Canada and the Administrative Department of Science, Technology and Innovation of Colombia (COLCIENCIAS)

Herein, special attention is given to whispered speech speaker verification (SV) as it is commonly used in public situations where a private or discrete information needs to be exchanged (e.g., when providing a credit card number, bank account number, or other personal information). Specifically, we address the mismatch whispered speech speaker verification problem, which is still an open issue specially during testing stage when there is no whispered speech data for training or to enroll target speakers [7], [8]. For instance, accuracies as low as 20% have been reported for whispered-speech speaker identification [7] in clean mismatch conditions.

In our previous research, different approaches were evaluated to tackle the mismatch problem including frequency warping and alternate feature representations, fusion schemes at frame and scoring level, and feature mapping [9], [10]. Other studies have proposed robust features such as modified linear cepstral coefficients (LFCC) and feature mapping [11], feature warping over Mel-frequency cepstral coefficients and score combination at the frame level [6], model adaptation schemes, and feature mapping from normal to whispered speech to be used during map adaptation [8]. Results suggested that there is invariant information between normal-voiced and whispered speech, but classical feature extraction approaches do not represent it in an effective manner to achieve reliable performance in a SV task for both speaking styles. Here, we propose the computation of innovative features aiming at extracting invariant information embedded within both speaking styles, by extracting variants of the MFCC. We present evidence on how the proposed approaches can extract complementary information to reduce the negative impact of the train/test mismatch problem.

The remainder of this paper is organized as follows. Section II presents background related to the SV problem. Section III describes the feature extraction approaches. Section IV describes databases used and discuses different approaches of fusion schemes. Section V, presents the results and discussion. Lastly, Section VI presents the conclusions.

II. Automatic speaker verification (SV)

Modern state-of-the-art speaker recognition systems are based on identity vectors extraction (*i-vectors*) [12], a technique proposed to map large-dimensional input data to a fixed-length low dimensional feature vector while retaining most

ISBN 978-0-9928626-7-1 © EURASIP 2017

relevant speaker information. The standard approach relies on the use of a C-Component Gaussian mixture model (GMM) trained as an universal background model (UBM) to partition the feature space and collect sufficient statistics. These statistics are, in turn, used to estimate a transformation matrix representing all variability in the space, the total variability matrix (T matrix), which is later used to extract i-vectors from speech recordings. In this approach, the number of components in the UBM model (C) and the dimensionality of the T matrix (D) need to be tuned. This procedure is complemented with some post-processing techniques such as linear discriminant analysis (LDA), whitening, and length normalization, which are helpful to remove nuisance effects in the total variability space. The interested reader is referred to [12], [13] for more complete details.

Matching between a test utterance and a target speaker can be done using either a fast scoring technique based on cosine distance or on probabilistic linear discriminant analysis (PLDA) [12], [14] based scoring. The PLDA model splits the total data variability into within-individual and between-individual variabilities, both residing in small-dimensional subspaces. Originally introduced for face recognition, PLDA has become a standard in speaker recognition, and details can be found in [14]. Using ϕ to denote the i-vector extracted from a given speech recording, in a verification scenario, there are two possible hypotheses: 1) ϕ_{test} and ϕ_{enrol} share the same class, and 2) ϕ_{test} and ϕ_{enrol} are from different classes. Lastly, the corresponding score can be obtained by computing the log-likelihood between the two hypotheses, which is given by $s = \ln(P(\phi_{test}, \phi_{enrol})) - \ln(P(\phi_{test})P(\phi_{enrol}))$; details can be found in [14]. For the experiments herein, the open-source *Bob* signal processing toolbox was used [15].

III. Feature extraction methods

The approach taken in this work is based on the computation of features aiming at extracting invariant information embedded within both speaking styles. First we use the standard mel-frequency cepstral coefficients in order to characterize a baseline system. Next, we complement this feature set with variants of MFCC that are proposed based on insights obtained from acoustic analyses.

A. Mel Frequency Cepstral Coefficients - MFCC

For the experiments herein, the MFCC were computed on a per-window basis using a 25 ms window with 40% overlap, the recordings were pre-emphasized using a first order finite impulse response filter with constant $a = 0.97$. Thirteen MFCC features, including the 0–th order cepstral coefficient (log-energy) were computed using 27 triangular bandpass filters spaced according to the mel scale. Delta (Δ) and double delta ($\Delta\Delta$) coefficients were appended to include dynamic or transitional information. These features were computed using an anti-symmetric Finite Impulse Response (FIR) filter of length nine to avoid phase distortion of the temporal sequence. After dropping frames where no vocal activity was detected, cepstral mean and variance normalization was applied per recording to remove linear channel effects. For all recordings the sampling rate is 16kHz.

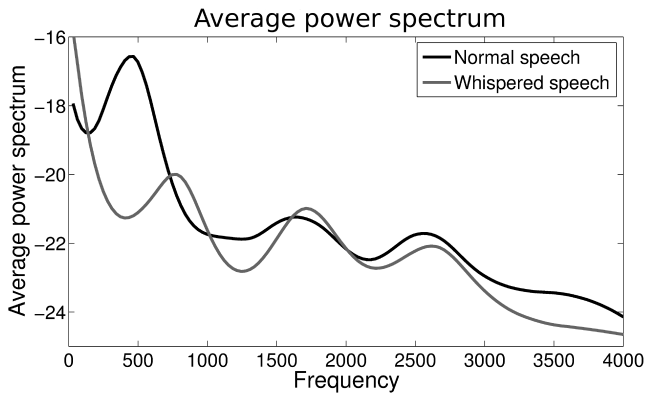

Fig. 1. Spectral envelope comparison between normal and whispered speech.

B. MFCC variants

According to perceptual and acoustic studies, two of the most salient differences between normal and whispered speech are related to the spectral envelope, i.e. *i*) whispered speech has a lower and flatter power spectral density [3] and *ii*) the formants shift towards higher frequencies [2], [16]. This last observation is more noticeable for the first three formants (F1, F2 and F3), where, F1 shifts can be up to 71% for men and 52% for women; F2 shifts can be up to 24% for men and 20% for women; and F3 shifts can be of 10% and 4.8%, respectively [16]. To illustrate this, Figure 1 depicts the average power spectrum of amplitude-normalized and pre-emphasized recordings from 36 speakers (male and female). As can be seen most of the differences remain below 1.2 kHz. For normal speech, most of the energy is concentrated below 1 kHz, whereas for whispered speech it is concentrated below 500 Hz, with frequency shifts in the spectral peaks and valleys (F1 shifts are most prominent). Based on these insights, two MFCC variants are proposed.

Based on the source-filter model for the process of human speech generation, it is possible to split the speech signal in two components, *i*) an excitation signal (known as the residual) that can be visualized as the combination of two different signal generators, one for voiced-speech and another for voiceless (noise-like) speech, and *ii*) a transfer function which models the vocal tract configuration and shapes the spectral envelope of the resulting speech [17]. This is relevant for whispered speech because by removing the influence of the vocal tract, then differences related to the spectral envelope are reduced. Furthermore, in the past, features extracted from the residual have been shown to contain important speaker-dependent information useful for speaker recognition tasks [18]. In the case of whispered speech, unvoiced sounds have been shown to remain unaffected [4], and also to contain important speaker-dependent information for speaker recognition tasks using whispered speech [19], [20]. As such, it is expected that features extracted from the residual signal will carry some invariant speaker information, particularly from the unvoiced segments. While it is not expected that the residual based feature will perform accurately alone, as most of the speaker-dependent information is typically embedded in spectral envelope associated to the vocal tract configuration, it should

ISBN 978-0-9928626-7-1 © EURASIP 2017

carry complementary information that can be fused with other features [21], [22]. For our experiments, the standard MFCC processing pipeline was computed over the residual signal. We will refer to this new feature set as Residual Mel Frequency Cepstral Coefficients (RMFCC). RMFCC, Δ and $\Delta\Delta$ are concatenated, then used for i-vector computation.

In the past, for narrowband (NB) speech signals, residuals were also explored for speaker verification of normal speech (i.e., 8kHz sampling) [22]. In such case, the residual modelled differences in excitation energy and periodicity information amongst speakers. Here, residuals are explored for alternate reasons, as there is no periodicity to be modelled with whispered speech due to the lack of vocal fold vibrations. Our hypothesis is that the resulting spectrally flat signal, even with the harmonic structure for normal speech, has reduced differences when comparing the two speaking styles. Furthermore, unvoiced sounds have more energy concentration at higher frequencies [17] and consonants such as stops, fricatives and affricates have more spectral similarities at frequencies higher than 4 kHz [3]. As such, by analyzing residuals from wideband (WB) speech, more information related to the unvoicedness will be captured. Given the similarities between the two vocal efforts for unvoiced speech segments, it is believed that this information will contain useful speaker-dependent information invariant across the two vocal efforts.

Typically, the use of NB signals for telephone based communications has limited the analysis of speech signals for feature extraction to the range of frequencies in 0.3 - 3.4 kHz. With the use of emerging WB communications and advanced digital signal processing technology in the telecommunications infrastructure, this range has been expanded to 8 kHz [23]. This has motivated detailed analyses to explore the role and relevance of different frequency subbands for speaker recognition tasks. As an example, for NB speech signals in [24] it was shown that the 1.5 - 3.4 kHz frequency sub-band contains more discriminative information than the lower 0.3 - 1.5 kHz frequency sub-band, except for nasals. For WB speech signals, on the other hand, in [25] it was shown that the frequency sub-band 4-8 kHz provides a performance similar to that obtained with the frequency sub-band 0-4 kHz, thus suggesting the presence of relevant speaker-discriminative information beyond 4 kHz. In a different study, it was shown that for text-dependent speaker identification, higher frequency channels were more relevant for speaker recognition than those located at lower frequencies [26]. It was reported that the lowest identification rates were associated to channels containing information of first and second formants, and that there was a high negative impact in performance when removing channels containing information from the frequency band between 5 kHz to 8 kHz [26]. Moreover, as seen from Fig. 1, the mismatch when comparing the sub-band from 0 to approximately 1.2 kHz is considerable, and in previous work it was shown that by removing this specific sub-band [9] it was possible to improve performance in the mismatch condition, but at the cost of reduced performance in the matched scenario. By doing this, we remove the sub-band that comprises mostly information from F1, which studies have shown to be highly variable in whispered speech mode, with shifts as high as 70% relative to normal speech. In addition to this, for WB speech signals it is expected that most of the speaker specific information relevant for speaker recognition tasks is preserved [26], [27], [28], hence the performance in normal speech should not be affected. This set is referred as Limited band Mel Frequency Cepstral Coefficients (LMFCC). As before, LMFCC, Δ and $\Delta\Delta$ are computed, then used for i-vector computation.

Database	No. of speakers		Recordings/speaker	
	Female	Male	Normal	Whisper
TIMIT	192	438	10	–
wTIMIT	24	24	450	450
CHAINS	16	20	37	37

TABLE I
DETAILS OF THE THREE DATABASES USED IN THIS WORK

IV. Experimental setup

A. Corpus description

In our experiments, three different databases were used, the CHAINS (Characterizing Individual Speakers) speech corpus [29], wTIMIT (whispered TIMIT) [30] and TIMIT databases [31]. The CHAINS and wTIMIT databases contain normal and whispered speech. Table I presents details about the number of speakers and recordings per speaker.

Speakers were divided in two disjoint sets, one for development (parameter estimation of GMM, T-matrix and PLDA) and the other for enrollment and testing (target speakers). Recordings from 462 speakers from TIMIT database and 14 speakers from wTIMIT were included in the development set, where the subset of speakers from wTIMIT includes recordings from both normal and whispered speech as suggested in [10]. Recordings from 100 speakers from the TIMIT database, 24 speakers from wTIMIT and 36 speakers from CHAINS, in turn, were included in the target speakers set. Average duration for all speech recordings is 4.5 seconds. It is important to emphasize that these are rather short utterances with limited phonetic variability thus making these experimental conditions more challenging [32]. To characterize the baseline system, we included only normal speech recordings from both the development and target speakers sets. During enrollment, eight recordings per speaker were used; for testing, however, we used two recordings per speaker, and if there are whispered speech recordings available, then two additional utterances were included. When combining the three databases the total number of target and non-target trials are 320 and 50880, respectively, for normal speech, and for whispered speech 120 target and 19080 non-target trials.

B. Fusion strategies

Three fusion schemes were also investigated in this paper: *i) score-level fusion*, *ii) i-vector concatenation* and *iii) Frame level fusion*. For score-level fusion, separate data (different from background and target speakers) is needed to train the fusion system and the systems to be fused (i.e., systems trained on MFCC, RMFCC or LMFCC feature sets) are evaluated using an unseen evaluation set. A logistic regression function is used as a fusion system and maps evaluation scores into a

ISBN 978-0-9928626-7-1 © EURASIP 2017

final decision using the Bosaris toolkit [33]. To estimate the parameters of this fusion system, 68 speakers from the TIMIT database and 10 from the wTIMIT database were included to create a new evaluation list. For enrollment, a configuration similar to the one used for the original evaluation list was used, including eight additional recordings of whispered speech for the 10 speakers of wTIMIT. For the new evaluation list, in order to have approximately the same amount of target and impostor scores from each speaking style, two recordings of normal speech and 15 recordings of whispered speech per speaker were used.

With i-vector fusion, in turn, i-vectors extracted from MFCC, RMFCC or LMFCC features are concatenated into a final feature vector prior to post-processing, i.e., prior to LDA, whitening, and length normalization. This strategy has shown to be effective in various scenarios such as language recognition and short utterance speaker recognition [34], [35] to combine strengths of i-vectors estimated from different feature representations. This approach does not require training of an additional system thus represents an advantage over score level fusion. Lastly, with frame level fusion, MFCC, RMFCC or LMFCC features (including first and second order derivatives) are concatenated into a final feature vector. Principal component analysis is then performed to remove redundant features and only the top components are kept as features, for combinations 99% of cumulative variance was retained. These top components are then used for i-vector computation.

V. Results and discussion

In previous work, we have shown the benefits of adding whispered speech during parameter estimation, even if the set of speakers with whispered speech recordings represents a small percentage of the overall training set [10]. Nevertheless, even if whispered speech data is available for parameter estimation, the mismatch problem can still be present as changes in the vocal effort can be viewed as "within-speaker" variation, and such variation is not well represented in the enrollment samples from target speakers [10]. To set the baseline of our experiments, we compared the performance of the three feature representations by training independent SV systems and adding whispered speech during T matrix estimation. Results are presented in Table II. As can be seen, by using the standard MFCC, there is a gap in performance between normal and whispered speech higher than 17%. Next, by using the RMFCC feature set, it is clear that by removing the information related to the spectral envelope important speaker specific information is removed. Systems based only on this feature set are not expected to perform at the same level as standard MFCCs, but may provide complementary information. And finally, LMFCC, are shown to perform equally well to MFCCs for normal speech, but to improve whispered speech accuracy by about 4%.

Table III, on the other hand, reports the results from the three fusion strategies. Best results are highlighted per speaking style. Results confirm that the three feature sets extract complementary and speaker-dependent discriminative information. It is necessary to highlight that the gains attained at this point, where only a small set of speakers in the background set have whispered speech recordings, shows that it is possible to extract invariant information from both speaking styles. Table III also shows four different combinations of feature sets, for example, S1 represents the fusion at different levels of MFCC and RMFCC feature representation. As can be seen, fusion of conventional MFCC with either of the proposed feature sets showed improvements for both normal and whispered speech (i.e., sets S1 and S2), relative to the results reported only for MFCC in Table II. Notwithstanding, fusion of only the proposed MFCC variants (i.e., set S3) resulted in further gains, particularly for whispered speech, with score-level fusion achieving the lowest EER for whispered speaking mode. Lastly, fusion of all features (i.e., set S4), while it did not improve the performance for whispered speech, it did slightly lower the EER for normal speech when using the frame-level fusion strategy. These results reinforce the idea that the proposed variants of MFCC extract more relevant speaker-dependent and complementary information which is invariant for both speaking styles than the standard MFCC.

Relative EER improvements, when comparing S3 with the standard MFCC based system, are 20% and 43% for normal and whispered speech, respectively, with score-level fusion. With i-vector concatenation, in turn, 30% and 36% gains for normal and whispered speech are seen, respectively. Lastly, frame-level fusion resulted in relative improvements of 37% for normal speech and 26% for whispered speech. By comparing the fusion schemes, it can be seen the one with best performance for whispered speech is score-level fusion, while the best for normal speech is frame-level fusion. i-vector concatenation, on the other hand, seems to be the scheme showing a tradeoff between performance and computational burden, as additional fusion scheme training is not needed, as was the case with score level fusion. This opens the possibility to implement dedicated systems for each speaking style, as different strategies showed benefits for different vocal efforts. We explored this alternative, and implemented a simple normal/whisper speech classification system using RMFCC i-vectors, which were the best at this task with a classification error = 0.45%. The resulting EER were 2.12% and 11.67%

SV system	Normal	Whispered
MFCC	3.13	20.83
RMFCC	6.70	22.95
LMFCC	3.13	16.67

TABLE II
Equal Error Rate (EER %) comparison for different feature sets and the fusion systems. LMFCC: limited band mel cepstral coefficients, RMFCC - residual mel cepstral coefficients. For these experiments $C = 256$ and $T = 400$

Fusion	S1		S2		S3		S4	
level	N	W	N	W	N	W	N	W
Score	2.81	15.63	2.81	15.83	2.50	**11.67**	2.50	13.49
i-vector	2.19	16.46	2.81	15.70	2.19	13.26	2.19	14.71
Frame	2.19	16.84	2.34	16.33	1.95	15.40	**1.88**	15.63

TABLE III
EER (%) comparison for fusion systems at three different levels. N: testing with normal speech, W: testing with whispered speech, S1: MFCC+RMFCC, S2: MFCC + LMFCC, S3: LMFCC + RMFCC, S4: MFCC+RMFCC+LMFCC. For all cases $C = 256$ and $T = 400$

ISBN 978-0-9928626-7-1 © EURASIP 2017

for normal and whispered speech in the SV task, respectively. Misclassification when selecting the best system, seems to affect more normal speech than whispered speech, however, error rates are inline with best results attained.

VI. Conclusions

This paper has addressed the issue of speaker verification (SV) based on whispered speech. Two variants of the conventional MFCC features were implemented in order to reduce error rates for SV with whispered speech while maintaining the performance with normal speech. The complementarity of features was then explored by means of three fusion schemes. According to our results, fusion schemes have different advantages depending on the testing speaking style, while for normal speech fusion at frame level seems to be optimal, for whispered speech it is better to train separate systems and fuse them at the score level. Dedicated systems seems to be an option to take the best of each configuration, however high performance must be guaranteed in the normal/whisper speech classification stage.

Regarding whispered speech speaker verification performance, when there is no whispered speech data available to enroll target speakers, whispered speech can induce severe detrimental effects on a standard MFCC/i-vector/PLDA based speaker recognition system. In fact, even if whispered speech data is available for parameter estimation, the mismatch problem can still be present as changes in the vocal effort can be viewed as "within-speaker" variation, and such variation is not well represented in the enrollment samples from target speakers. Proposed schemes can partially address this problem by reducing the within-speaker variation, but it is necessary to continue exploring alternate feature representations to extract more relevant speaker-dependent invariant information across vocal efforts. In the future, feature selection methods can be used in order to reduce computational burden.

References

[1] V.C. Tartter, "Identifiability of vowels and speakers from whispered syllables," *Perception & Psychophysics*, vol. 49, no. 4, pp. 365–372, April 1991.

[2] M. Higashikawa, K. Nakai, A. Sakakura, and H. Takahashi, "Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study," *J. Voice*, vol. 10, no. 2, pp. 155–158, 1996.

[3] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, February 2005.

[4] S.T. Jovicic and Z. Saric, "Acoustic analysis of consonants in whispered speech," *Journal of Voice*, vol. 22, no. 3, pp. 263–274, May 2008.

[5] H.R. Sharifzadeh, I.V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, October 2010.

[6] Q. Jin, S.-C. Jou, and T. Schultz, "Whispering speaker identification," in *IEEE International Conference on Multimedia and Expo*, July 2007, pp. 1027–1030.

[7] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1097–1111, August 2008.

[8] X. Fan and J.H.L. Hansen, "Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams," *Speech Communication*, vol. 55, no. 1, pp. 119–134, January 2013.

[9] M. Sarria-Paja and T.H. Falk, "Strategies to enhance whispered speech speaker verification: A comparative analysis," *Journal of the Canadian Acoustical Association*, vol. 43, no. 4, pp. 31–45, 2015.

[10] M. Sarria-Paja, M. Senoussaoui, D. O'Shaughnessy, and T.H. Falk, "Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification," in *Proc. ICASSP*, March 2016, pp. 5480–5484.

[11] X. Fan and J. H. L. Hansen, "Speaker identification with whispered speech based on modified LFCC parameters and feature mapping," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 4553–4556.

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

[13] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *Proc. INTERSPEECH*, 2011, pp. 485–488.

[14] A Sizov, K.A. Lee, and T. Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," in *Proc. S+SSPR*, 2014.

[15] A. Anjos et al., "Bob: a free signal processing and machine learning toolbox for researchers," in *Proc. 20th ACM Conference on Multimedia Systems*, Oct. 2012.

[16] H.R. Sharifzadeh, I.V. McLoughlin, and M. Russell, "A comprehensive vowel space for whispered speech," *Journal of Voice*, vol. 26, no. 2, pp. 49–56, March 2012.

[17] D. O'Shaughnessy, *Speech communications - human and machine (2. ed.)*, IEEE, 2000.

[18] T. Drugman and T. Dutoit, "On the potential of glottal signatures for speaker recognition," in *Proc. INTERSPEECH*. 2010, pp. 2106–2109, ISCA.

[19] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1408–1421, July 2011.

[20] J. Xu and H. Zhao, "Speaker identification with whispered speech using unvoiced-consonant phonemes," in *Proc. IASP*, Nov 2012, pp. 1–4.

[21] Md. Sahidullah, S. Chakroborty, and G. Saha, "Improving performance of speaker identification system using complementary information fusion," *CoRR*, vol. abs/1105.2770, 2011.

[22] P. Debadatta and M. Prasanna, "Processing of linear prediction residual in spectral and cepstral domains for speaker information," *International Journal of Speech Technology*, vol. 18, no. 3, pp. 333–350, 2015.

[23] J. Rodman, "The effect of bandwidth on speech intelligibility - white paper," Tech. Rep., POLYCOM Inc., USA, September 2006.

[24] H. Lei and E. Lopez-Gonzalo, "Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition," in *Proc. INTERSPEECH*, 2009.

[25] L. Gallardo, M. Wagner, and S. Möller, "Advantages of wideband over narrowband channels for speaker verification employing MFCCs and LFCCs.," in *Proc. INTERSPEECH*, 2014.

[26] L. Besacier and J.-F. Bonastre, "Subband approach for automatic speaker recognition: Optimal division of the frequency domain," in *Proc. First International Conference on Audio- and Video-based Biometric Person Authentication*, March 1997, Lecture Notes in Computer Science, pp. 195–202.

[27] K. McDougall, "Speaker-specific formant dynamics: An experiment on australian english /aI/," *Speech, Language and the Law.*, vol. 11, no. 1, pp. 103–130, June 2004.

[28] K. McDougall and F. Nolan, "Discrimination of speakers using the formant dynamics of /u:/ in british english," in *Proc. 16th International Congress of Phonetic Sciences*, August 2007, pp. 1825–1828.

[29] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: Characterizing individual speakers," in *Proc of SPECOM*, 2006, vol. 6, pp. 431–435.

[30] Boon Pang Lim, *Computational differences between whispered and non-whispered speech*, Ph.D. thesis, University of Illinois, 2011.

[31] J.S. Garofolo, Linguistic Data Consortium, et al., "TIMIT: acoustic-phonetic continuous speech corpus," 1993.

[32] A. Kanagasundaram, R. Vogt, D.B. Dean, S. Sridharan, and M.W. Mason, "I-vector based speaker recognition on short utterances," in *Proc. INTERSPEECH*, 2011, pp. 2341–2344.

[33] N. Brummer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, algorithms and code for binary classifier score processing," Tech. Rep., CAGNITIO Research, South Africa, 2011.

[34] S. Irtza, H. Bavattichalil, V. Sethu, and E. Ambikairajah, "Scalable i-vector concatenation for plda based language identification system," in *Proc. APSIPA*, Dec 2015, pp. 1182–1185.

[35] Z.-Y. Li, W.-Q. Zhang, and J. Liu, "Multi-resolution time frequency feature and complementary combination for short utterance speaker recognition," *Multimedia Tools and Applications*, vol. 74, no. 3, pp. 937–953, 2015.

ISBN 978-0-9928626-7-1 © EURASIP 2017