IMPOSTURE CLASSIFICATION FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li

IMPOSTURE CLASSIFICATION FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li,

Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore

{*alarcher,kalee,mabin,hli*}*@i2r.a-star.edu.sg*

ABSTRACT

This work focuses on text-dependent speaker verification, where a user is required to chose and pronounce a customized pass-phrase to get authenticated. In this context, there are three types of impostures: an impostor pronouncing the correct pass-phrase, an impostor pronouncing a wrong pass-phrase and the most difficult one: an impostor playing back a recording of the target speaker pronouncing a wrong pass-phrase. Detecting and classifying different types of impostures can help to prevent future impostures of the same type. In this work, we first propose a new verification score to reject *Playback* impostures. This score allows a relative reduction of 90% of the equal error rate against *Playback* impostures while offering performance similar to the baseline text-dependent score against other types of impostures. As a second contribution, we show that the new score can be combined with an existing text-dependent verification score to improve the classification of the different types of impostures. The performance of the speaker verification engine for imposture classification is significantly improved with the C_{llr} decreasing by at least 29% compared to the original system.

***Index Terms*—** Speaker verification, Text-Dependent, Impostures, Playback

1. INTRODUCTION

Speaker verification is the task of accepting or rejecting an identity claim based on the information extracted from a voice sample [1]. This task consists of classifying two types of trials: the genuine trials that have to be accepted and the impostures trials that must be rejected. This work deals with a specific scenario of text-dependent speaker verification [2], where the customer is required to chose a personal pass-phrase and to pronounce it to be authenticated. In this context, three types of impostures can be defined given the the nature of the speaker (target or impostor), and the pass-phrase pronounced (correct or wrong).

Amongst the possible impostures, illustrated in Fig.1, the *Naive* imposture consists of an impostor pronouncing a pass-phrase different from the one chosen by the user (s)he is imposting. This imposture where both the identity and the pass-phrase are different from the ones expected by the system is supposed to be easily rejected. On the contrary, the two other types of imposture represent more serious threats to the system as either the pass-phrase or the speaker identity is correct. In this work, we refer to the case of an impostor pronouncing the correct pass-phrase as *Sly* imposture, while *Playback* imposture will be use for the case of the target speaker pronouncing a sentence different from his/her personal pass-phrase. This last case refers to the situation where an impostor playbacks a recorded voice of the target speaker in order to spoof the system.

This work aims at increasing the performance of a single speaker verification engine to discriminate between the different types of impostures. Indeed, *Naive* impostures are relatively easy to reject and the cost of accepting this imposture may not be high if we consider that the impostor did not especially prepared his/her attack to the system. On the contrary, a *Sly* or *Playback* imposture requires additional preparation efforts and might involves malevolent intention. It might also imply that the impostor will attack again in the future and knowledge about imposture attempts would help preserve the integrity of the system. For instance, when detecting a *Sly* imposture, the system can ask the speaker to change the pass-phrase, already known by the impostor. For a *Playback* imposture, the system can keep a copy of the speech sample so as to detect future use of this sample or to identify the impostor.

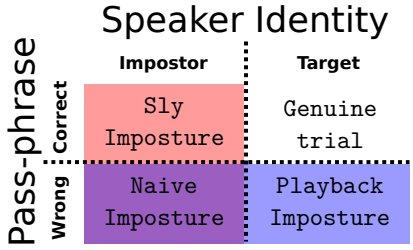

Fig. 1. Text-dependent speaker verification systems encounter 4 types of trials whether the speaker is the target speaker or an impostor who pronounces the correct pass-phrase or a wrong one.

Prior works considering *Playback* impostures propose to make use of a speech recognition system to reject a user pronouncing a wrong pass-phrase [3] or to use a second biometric modality to reject impostors playing back a recording [4]. However, these solutions imply an extra computational cost due to the additional system, that is not suitable for all applications [5]. Multi-modality can also be used to thwart *Sly* impostures [6] but lighter approaches involving only speech processing have been proposed. In [7], a HMM-based system is reinforced by an additional duration information. This method is not suitable for the case of user-customized pass-phrases as it requires to train a speaker-independent HMM model for each pass-phrase. Another approach includes high-level information such as pitch contour or source information in a dynamic programming framework [8]. However, this approach adds in complexity and might be more sensitive to session variability.

In this work we propose to tackle different types of impostures by using a single engine, thus not increasing the computational cost of the verification. Based on the existing HiLAM engine [9, 10, 11], we first propose a new score that shows better discrimination against *Playback* and *Naive* impostures while offering similar performance against the *Sly* impostures. We then show that combining this new score with the one originally proposed in [9] into a dual scoring allows a better separation of the three different types of impostures.

2. MODELING BACKGROUND, SPEAKER AND TEXT

Classical text-independent speaker verification engines are based on the GMM/UBM paradigm [12, 13]. Given a sequence of features

$\mathcal{X}$ at testing time, the verification score produced is a log-likelihood ratio between the hypothesis, H_0, that the speech segment has been spoken by the target speaker and its alternative hypothesis, H_1, that is was spoken by an impostor. A speaker and text independent Universal Background Model (UBM), trained on a relatively large amount of data, is used to model hypothesis H_1 while a speaker-dependent, text-independent Gaussian mixture model (GMM) is adapted from the UBM by using all data available from the target speaker. The verification score is given by:

$$S_{ti}(\mathcal{X}) = \log \frac{L_{\lambda_{gmm}}(\mathcal{X})}{L_{\lambda_{ubm}}(\mathcal{X})} \quad (1)$$

where $L_{\lambda_{gmm}}(\mathcal{X})$ and $L_{\lambda_{ubm}}(\mathcal{X})$ are respectively likelihood of $\mathcal{X}$ over the speaker's text-independent GMM and the UBM. However, the performance of such system degrades strongly when the duration of speech material is limited.

Text-dependency is well known to compensate for the lack of speech material by harnessing the temporal structure and phonetic content of a specific pass-phrase [2, 14]. Modeling of both the speaker and the pass-phrase is commonly done using Hidden Markov Models (HMMs) which offer a relative robustness to speaker and environment variabilities. In this context, the computation of a likelihood ratio for speaker verification often makes use of a speaker-independent HMM to model the alternative hypothesis H_1 [15, 16, 17]. This approach is not suitable for the case of user-customized pass-phrase and we proposed in a previous [9, 10, 11] work to derive the pass-phrase-dependent HMM from a UBM that is also used to model the alternative hypothesis for any chosen pass-phrase.

In our previous work [10, 11], a text-dependent log-likelihood ratio, $S_{td}(\mathcal{X})$, is computed as:

$$S_{td}(\mathcal{X}) = \log \frac{L_{\lambda_{hmm}}(\mathcal{X})}{L_{\lambda_{ubm}}(\mathcal{X})} \quad (2)$$

where $L_{\lambda_{hmm}}(\mathcal{X})$ is the likelihood of $\mathcal{X}$ over the speaker's text-dependent HMM aligned by Viterbi decoding.

The resulting architecture, called HiLAM, has been recently deployed in a large scale commercial application [5]. A complete description of HiLAM and its training process can be found in [9, 10, 11]. Considering the aim of this work, essential information regarding the HiLAM architecture can be summarized as follows:

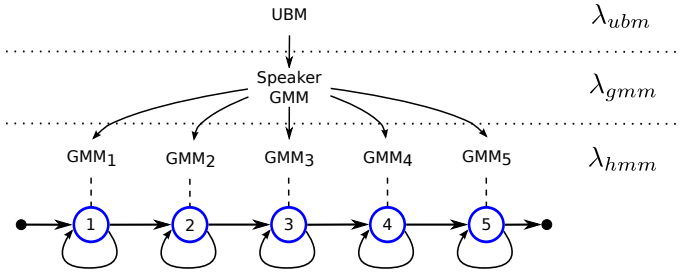

Fig. 2. The Hierarchical multi-Layer Acoustic Model (HiLAM). The first two layers are similar to the standard GMM/UBM while the bottom layer hinges on the abilities of a left-right HMM to harness the specific temporal structure of pass-phrases.

- the first layer is the classical, speaker- and text-independent, Universal Background Model (UBM)
- the middle layer is a speaker-dependent and text-independent GMM with its means adapted from the first layer UBM
- the bottom layer is a speaker- and text-dependent HMM modeling the user-specific pass-phrase. All state's density distributions of this HMM are GMMs with its mean adapted from the middle layer speaker model
- all nodes in the HiLAM architecture are N-distribution GMMs sharing the same covariance and weight coefficients

The acoustic features used in our experiments are 50-dimension vectors composed of 19 MFCC, their derivatives, 11 first second derivatives and delta energy. Feature frames are computed on a 20ms sliding window with shifting of 10ms. Low-energy frames are discarded and mean-variance normalization is applied. Each node of the HiLAM architecture is a GMM with N=64 mixtures.

3. IMPOSTURE CLASSIFICATION

3.1. Playback detection

Playbacks are the most difficult impostures to reject. For both text-independent (S_{ti}) and -dependent scores (S_{td}), the denominator of the likelihood ratio is computed by using the UBM which is assumed to model all speakers except the target. Therefore, none of this score is designed to reject *Playback* impostures involving recording from the target speaker. In order to explicitly tackle the *Playback* impostures, we propose to use a new score based on the log-likelihood ratio of $\mathcal{X}$ over the speaker's text-dependent HMM aligned by Viterbi decoding and the speaker text-independent GMM such that:

$$S_{sn}(\mathcal{X}) = \log \frac{L_{\lambda_{hmm}}(\mathcal{X})}{L_{\lambda_{gmm}}(\mathcal{X})} \quad (3)$$

This score, hereafter referred to as *Speaker Normalized*, S_{sn}, explicitly compares the hypothesis of the target speaker pronouncing the correct pass-phrase, modeled by a speaker- and text-dependent HMM, to the hypothesis of the target speaker pronouncing any pass-phrase, modeled by a speaker-dependent and text-independent GMM. Note that this score can be expressed as the difference between the text-dependent and text-independent scores: $S_{sn}(\mathcal{X}) = S_{td}(\mathcal{X}) - S_{ti}(\mathcal{X})$. A direct comparison of these two scores is given by Figure 3 and provides more insight. The use of the additional text-dependent component of the HiLAM does not affect the score distributions of *Naive* and *Playback* impostures where the impostor pronounces a wrong pass-phrase (Figures 3-B and -D). On the contrary, the text-dependent component of the HiLAM architecture increases the scores of speaker pronouncing the correct pass-phrase. This increase, very limited for the *Sly* impostures (Figure 3-C), is very significant for the *Genuine* trials (Figure 3-A).

3.2. Dual scoring for imposture classification

We propose now to classify the different types of impostures in a two dimensional space obtained by combining the the original text-dependent score, S_{td}, and the proposed *Speaker Normalized* score, S_{sn}, introduced in Section 3.1. The different behaviors exhibited by these two scores is expected to improve the separation of the four types of trials. The benefit of the resulting dual-score, of the form $\boldsymbol{s} = [S_{td}; S_{sn}]^T$, is illustrated by the Figure 4 that shows the four types of trial in this two-dimensional space. Note that the repartition of the four types of trials on this figure presents a certain similarity with the theoretical framework represented in Figure 1.

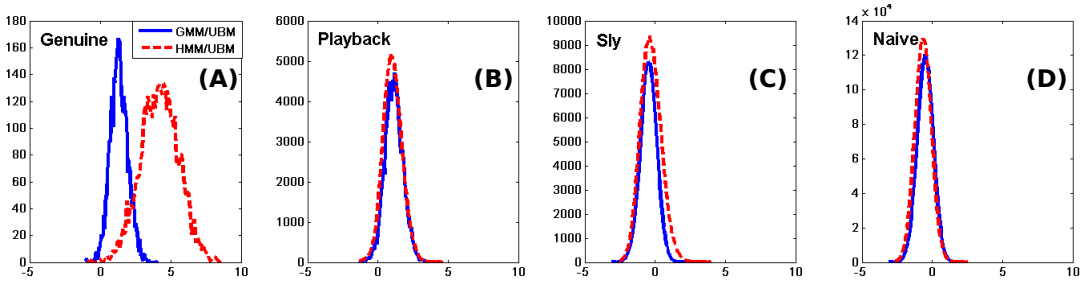

Fig. 3. Distribution of the GMM/UBM and HMM/UBM likelihood ratio scores generated by the HiLAM system for four different types of trials. The addition of the temporal information in the in the HMM/UBM score significantly increases the scores of the *Genuine* trials while keeping the distributions of *Playback*, *Sly* and *Naive* impostures unchanged.

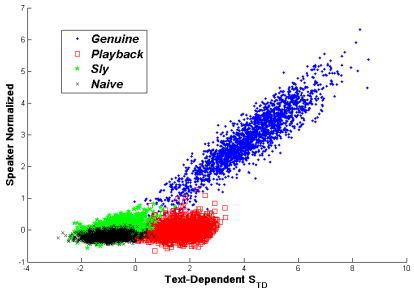

Fig. 4. Representation of a subset of scores for the four types of trials: *Genuine*, *Playback*, *Sly* and *Naive*, in a 2-dimensional space. The two dimensions of the space correspond to the text-dependent score, S_{td} and the *Speaker Normalized* score, S_{sn}.

Classification in the dual-score space can be done using classifiers such as Multi-class Logistic Regression [18] or multi-class SVM [19]. In order to demonstrate the potential of the dual-scoring, a gender-dependent hetero-scedastic Gaussian back-end[1] is trained on the scores of a development set. Given the collection of output score vectors computed on the development set, a multivariate normal distribution is trained for each type of trial by using the Maximum Likelihood criteria, as described in [20]. During the test, given an output score vector, $\boldsymbol{s}$, the final classification score for the corresponding trial is obtained by evaluating the log-likelihood of the vector $\boldsymbol{s}$ over the Gaussian distributions learned for each type of trial.

4. DATA, PROTOCOL AND EVALUATION METRICS

A text-dependent database, different from its text-independent counterpart, requires the recording of the same set of sentences across different speakers. To this end, Part I of the *RSR2015* database meets this requirement with sufficiently large number of speakers though the channel effects are benign [10, 11]. The 300 speakers of the *RSR2015* database are divided into three non-overlapping groups referred to as *background*, *development* and *evaluation*. Each speaker pronounces a set of 30 fixed pass-phrases across 9 sessions. To avoid the use of the 30 pass-phrases in background training, two gender-dependent UBMs are trained by using different lexical material from the Part II and III material of the *background* data.

Performance of the HiLAM system is evaluated on the *development* and validated on the *evaluation* set of the Part I of the *RSR2015* database. Out of the 9 sessions, 3 are used for enrollment and 6 for test. During the enrollment, three occurrences of the first 15 pass-phrases of a speaker are used to adapt a speaker-dependent GMM. For each of these 15 pass-phrases, a text-dependent HMM is adapted from this GMM to produce 15 text-dependent models of a same speaker. During the test, all 30 sentences from the 6 test sessions of a speaker are used to generate *Genuine* target trials and *Playback* impostures. The *Sly* and *Naive* impostures are generated by comparing all 30 sentences from the 6 test sessions of a speaker against all the models trained for the remaining speakers of the group, i.e. *development* or *evaluation*. Note that we don't consider cross gender tests and that 15 pass-phrases out of the 30 have been used during the enrollment phase while the 15 others have never been seen by the system. The total number of tests for each type of trial is given in Table 1. In the remaining of this paper, speaker verification performance is reported in terms of Equal Error Rate (EER) for each type of imposture separately. The deployment of an engine classifying the different types of impostures would require to fix the cost of the misclassification errors for evaluating the classifier performance. In order to eschews the use of specific mis-classification costs, the discriminancy is evaluated by using a log-likelihood-ratio-based performance measure, the multi-class C_{llr} [21]:

Table 1. Number of tests per speaker-set for each type of trial.

Type of trial	Male		Female	
	dev	eval	dev	eval
Genuine	4,479	5,089	4,199	4,303
Playback	129,906	147,611	121,771	124,832
Sly	219,472	284,987	193,157	202,242
Naive	3,185,841	4,133,183	2,799,411	2,933,184

$$C_{llr} = -\frac{1}{T}\sum_{t=1}^{T} w_t \log_2 P_t \quad (4)$$

where P_t is the posterior probability of the true class of the trial t calculated for a flat prior and w_t is a weighting factor that normalized the class proportions in the test set. Multi-class C_{llr} is a positive value expressed in bits of information that measures the actual performance of the classifier (the lower the better). For comparison, C_{llr}^{min} and *Reference Loss* are also provided. C_{llr}^{min} is the C_{llr} value

[1] https://sites.google.com/site/nikobrummer/focalmulticlass

obtained for an optimum calibration of the scores; it reflects the potential of the classifier without considering the calibration issue. The *Reference Loss* is the C_{llr} value of a system that extract no information from the speech signal for the given task.

5. EXPERIMENTS

The first experiment is conducted on the *RSR2015* development set to compare the performance of the three scores introduced previously. Even though the system performs better for female speakers, the behavior of the different scores is consistent across gender. The GMM/UBM text-independent score, S_{ti}, which is given as reference, obtains the worse performance for all impostures (Table 2). Especially, the text-independent score is not good to reject the *Playback* impostures due to its lack of lexical information. *Naive* impostures, however, are better rejected than *Sly* impostures, probably because the training material of the speaker text-independent GMM covers the lexical content of half of the test pass-phrases. It shows the speaker GMM not to be completely text-independent while trained with 15 different pass-phrases.

Table 2. Performance of the HiLAM system on the *development* part of the *RSR2015* database. Performance is given in terms of EER (%) for three scores: the text-independent GMM/UBM score, S_{ti}, the text-dependent score, S_{td} and the *Speaker normalized* score, S_{sn}.

Imposture	Male			Female		
	S_{ti}	S_{td}	S_{sn}	S_{ti}	S_{td}	S_{sn}
Playback	43.48	6.23	0.59	42.99	2.50	0.22
Sly	6.14	1.82	1.90	5.29	0.93	0.88
Naive	5.53	0.59	0.20	4.63	0.12	0.07

The text-dependent score, S_{td}, greatly reduces the error rates for all impostures (columns 3 and 6 of Table 2). Nevertheless, EER remains at 6.23% for male and 2.50% for the female for *Playback* impostures, which are still the most difficult to reject. During enrollment, 15 of the 30 available pass-phrases have been used to train the speaker models. As a consequence, *Playback* impostures using one of these 15 pass-phrases are more difficult to reject than the 15 unseen pass-phrases. EER obtained for both seen and unseen sentences respectively vary from 7.21% to 5.27% for male and from 2.95% to 1.91% for female speakers.

The proposed S_{sn} is expected to thwart the three types of impostures together and especially the *Playback*. *Playback* impostures are better rejected than when using the two other scores (columns 4 and 7 of Table 2). Compared to the original text-dependent score, the EER reduces from more than 90% for both male and female speakers, reaching respectively 0.59% and 0.22%. *Playback* impostures generated with the 15 pass-phrases seen during the enrollment are now better rejected than the 15 unseen pass-phrases, due to the fact that the GMM speaker model used for the denominator of the likelihood ratio was trained using these pass-phrases. EER obtained for both seen and unseen sentences now vary from 0.38% to 0.67% for male and from 0.14% to 0.24% for female speakers. We can see that the *Speaker normalized* score also provides improvement against *Naive* impostures and preserves performance of the text-dependent score, S_{td}, against *Sly* impostures.

In a second experiment, we evaluate the ability of the system to discriminate between the four types of trials (*Genuine*, *Playback*, *Sly* and *Naive*). Classification performance using the dual-score is evaluated in terms of C_{llr} and C_{llr}^{min} and compared to a baseline for which the input of the Gaussian back-end is a mono-dimensional score consisting of the original text-dependent score: S_{td} alone. This simple baseline is motivated by the lack of equivalent approach in the literature. This baseline present also the advantage to be similar to our approach in terms of complexity as the computational cost of the likelihood of the test utterance over the second layer GMM and of the Gaussian back-end is negligible. Performance of the two back-ends are given in terms of C_{llr} and C_{llr}^{min} in Table 3.

Table 3. Performance of the dual-score compared to the HiLAM original text-dependent score on the development and evaluation parts of the *RSR2015* database. Performance is given in terms of C_{llr} and C_{llr}^{min}. Reference loss is given for comparison

		Male		Female	
		Dev	*Eval*	*Dev*	*Eval*
Text-dependent	C_{llr}	0.9071	0.9429	0.8297	0.8860
	C_{llr}^{min}	0.9069	0.9410	0.8271	0.8774
Dual-Score	C_{llr}	0.5941	0.6110	0.6055	06325
	C_{llr}^{min}	0.5896	0.6075	0.5857	0.6061
Reference loss		2			

Compared to the baseline, the dual-score strongly reduces the C_{llr} for both genders. Indeed, the C_{llr} is reduced by a relative 35% for male speakers and 29% for the female speakers. Performance on the development set are given as reference as the Gaussian back-end has been trained on these data. However, we can see that the improvement on development and evaluation set is consistent.

6. DISCUSSION

The first contribution of this work is the detection of playback impostures. This is accomplished by combining a speaker-dependent text-independent GMM with a HMM which is both speaker and text-dependent. In our experiments, it was found that this new score greatly outperforms the existing text-dependent score when rejecting *Playback* impostures for which the target speaker pronounces a wrong pass-phrase. The EER obtained against this type of imposture is reduced by 90% for both male and female speakers compared to the original text-dependent score. Additionally, the S_{sn} score also outperforms the baseline text-dependent score when considering the case of an impostor pronouncing a wrong pass-phrase.

The second contribution of this work consists of combining two verification scores, the original text-dependent score and the new S_{sn} score, into a dual-score to improve the classification of the four types of trials encountered by a text-dependent speaker verification system. Integrated into a Gaussian back-end, the dual-score provides a significant improvement compared to the baseline as the C_{llr} decreases by at least 29% for both male and female speakers. This improvement is obtained without significant increase of the computational cost as both scores are obtained from the same HiLAM architecture.

Improving the rejection of the *Playbacks* and the classification of the different types of impostures is expected to benefit to the security of text-dependent speaker verification systems by allowing the prevention of future impostures. For further work we aim to increase the flexibility of this approach and provide similar improvements in different text-dependent scenarios such as scenarios using prompted text for liveness detection.

7. REFERENCES

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

[2] Matthieu Hébert, *Text-dependent speaker recognition*, Springer-Verlag, Heidelberg, 2008.

[3] Larry Heck and Dominique Genoud, "Integrating Speaker and Speech Recognizers: Automatic Identity Claim Capture for Speaker Verification," in *Odyssey Speaker and Language Recognition Workshop*, 2001, pp. 249–254.

[4] Girija Chetty and Michael Wagner, "Liveness detection using cross-modal correlations in face-voice person authentication," in *European Conference on Speech Communication and Technology (Eurospeech)*, 2005, pp. 2181–2184.

[5] Kong Aik Lee, Bin Ma, and Haizhou Li, "Speaker verification makes its debut in smartphone," in *SLTC Newsletter*, February 2013.

[6] Elie Khoury, Manuel Günther, Laurent El Shafey, and Sébastien Marcel, "On the Improvements of Uni-modal and Bi-modal Fusions of Speaker and Face Recognition for Mobile Biometrics," in *Biometric Technologies in Forensic Science*, 2013.

[7] Nestor Becerra Yoma and Tarciano Facco Pegoraro, "Robust speaker verification with state duration modeling," *Speech Communication*, vol. 38, no. 1–2, pp. 77–88, 2002.

[8] B Yegnanarayana, SR Mahadeva Prasanna, Jinu Mariam Zachariah, and Cheedella S Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 575–582, 2005.

[9] Kong Aik Lee, Anthony Larcher, Helen Thai, Bin Ma, and Haizhou Li, "Joint Application of Speech and Speaker Recognition for Automation and Security in Smart Home," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 3317–3318.

[10] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "The RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2012, pp. 1580–1583.

[11] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent Speaker Verification: Classifiers, Databases and RSR2015," *Speech Communication*, 2014, Accepted for publication.

[12] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[13] Frederic Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meigner, Teva Merlin, Javier Ortega-Garcia, Dijana Petrovska-Delacretaz, and Douglas A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, April 2004.

[14] Anthony Larcher, Jean-Francois Bonastre, and John S.D. Mason, "Reinforced temporal structure of acoustic models for speaker recognition," *Digital Signal Processing*, vol. 23, no. 6, pp. 1910–1917, December 2013.

[15] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing]*, vol. 29, no. 2, pp. 254–272, 1981.

[16] Delphine Charlet, Denis Jouvet, and O. Collin, "An alternative normalization scheme in HMM-based text-dependent speaker verification," *Speech Communication*, vol. 31, no. 2-3, pp. 113–120, 2000.

[17] Mohamed Faouzi BenZeghiba and Hervé Bourlard, "User-customized password speaker verification using multiple reference and background models," *Speech Communication*, vol. 48, no. 9, pp. 1200–1213, 2006.

[18] David A. Van Leeuwen and Niko Brümmer, "Channel-dependent GMM and Multi-class Logistic Regression models for language recognition," in *Odyssey Speaker and Language Recognition Workshop*, 2006.

[19] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks,*, vol. 13, no. 2, pp. 415–425, 2002.

[20] Haizhou Li, Bin Ma, and Kong Aik Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, 2013.

[21] Niko Brümmer and David van Leeuwen, "On calibration of language recognition scores," in *Odyssey Speaker and Language Recognition Workshop*. IEEE, 2006, pp. 1–8.