INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Utterance Verification for Text-Dependent Speaker Recognition: a Comparative Assessment Using the RedDots Corpus

Tomi Kinnunen[1], *Md Sahidullah*[1], *Ivan Kukanov*[1], *Héctor Delgado*[2], *Massimiliano Todisco*[2]
Achintya Sarkar[3], *Nicolai Bæk Thomsen*[3], *Ville Hautamäki*[1], *Nicholas Evans*[2], *Zheng-Hua Tan*[3]

[1]Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Finland
[2]Digital Security, EURECOM, France
[3]Signal and Information Processing, Department of Electronic Systems, Aalborg University, Denmark
tkinnu@cs.joensuu.fi, evans@eurecom.fr, zt@es.aau.dk

Abstract

Text-dependent automatic speaker verification naturally calls for the simultaneous verification of speaker identity and spoken content. These two tasks can be achieved with automatic speaker verification (ASV) and utterance verification (UV) technologies. While both have been addressed previously in the literature, a treatment of simultaneous speaker and utterance verification with a modern, standard database is so far lacking. This is despite the burgeoning demand for voice biometrics in a plethora of practical security applications. With the goal of improving overall verification performance, this paper reports different strategies for simultaneous ASV and UV in the context of short-duration, text-dependent speaker verification. Experiments performed on the recently released RedDots corpus are reported for three different ASV systems and four different UV systems. Results show that the combination of utterance verification with automatic speaker verification is (almost) universally beneficial with significant performance improvements being observed.

Index Terms: Speaker recognition, Text-dependent, Utterance verification, Confidence measure.

1. Introduction

Over the past two decades research in automatic speaker verification (ASV) [1, 2] has been driven largely by the speaker recognition evaluation (SRE) benchmarks organised by the National Institute of Standards and Technology (NIST) [3] in the US. This work has focused exclusively on text-*independent* tasks. Recent progress in the area is detailed in [4] and [5].

The predominant applications for text-independent ASV involve surveillance and forensics. In contrast, text-*dependent* ASV has utility in a wide range of user authentication applications, for example smart-phone log-in [6], telephone banking and physical access control. These scenarios typically demand the use of convenient, short pass-phrases. The quest for reliable performance then dictates strict text constraints, i.e., the same text for enrolment and authentication.

A text-dependent ASV system might use a fixed pass-phrase for all users. Alternatively, a different pass-phrase can be assigned to, or selected by each user individually. A third option, which can also help to protect from replay spoofing attacks [7, 8], involves the use of randomly prompted pass-phrases. In these cases, successful authentication would require the recognition of not just the speaker, but also the pass-phrase. This calls for the combination of ASV with some form of utterance verification (UV). This is the focus of this paper.

Two forms of UV are possible. The first is an implicit approach whereby UV is performed in unison with ASV, for instance with a hidden Markov Model (HMM) approach to speaker recognition. The second is an explicit approach in which ASV and UV are applied separately. In this case UV might be implemented as an auxiliary classifier tasked with accepting or rejecting the hypothesis that a given utterance contains an expected phone sequence.

Utterance verification can be viewed as *verbal information verification* or *spoken-content verification*. Even though it could be integrated with speaker recognition system to improve the security in practical scenario (e.g. prompting a text to avoid replay attack), this has not been studied much, possibly because the focus was given to text-independent ASV research. In limited research work available in this field, two systems are combined by fusing scores to find a global threshold for decision making as well as by a two-stage process with separate decision thresholds for UV and ASV [9, 10, 11].

Since speaker and phone variation are different entities, this paper investigates the second approach. Since this involves the combination of two classifiers, the manner in which UV and ASV should best be integrated is an open question. The paper investigates different approaches to UV in addition to strategies for its integration. Central to the work is an emphasis on convenient, short pass-phrases and improved overall verification performance.

2. Problem definitions

For a self-contained exposition, we define here the three tasks addressed in this study: stand-alone speaker verification, stand-alone utterance verification, and their combination.

2.1. Automatic speaker verification (ASV)

Let $\mathcal{U}$ be a speech utterance, represented using a collection of feature vectors $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. In **speaker verification**, given a claim of speaker identity, say speaker $j \in \{1, \ldots, J\}$, the goal is to evaluate a log-likelihood ratio score,

$$\ell_{\text{spk}}(\mathcal{X}, j) = \log \frac{p(\mathcal{X}|\text{same speaker})}{p(\mathcal{X}|\text{different speaker})}, \qquad (1)$$

where the hypotheses in the numerator and denominator are evaluated, respectively, using an adapted target speaker GMM and a universal background model (UBM) [12]. Alternatively, one can use i-vectors [13] as input features and evaluate (1) using, for instance, probabilistic linear discriminant

http://dx.doi.org/10.21437/Interspeech.2016-1125

analysis (PLDA) [14] scoring. To support dynamically increasing speaker databases along the life-cycle of an ASV system, the anti-model training data usually originates from speakers disjoint from any of the targets. More detailed treatments of speaker verification technology are available in [4, 5] (text-independent) and [15] (text-dependent). Details of our ASV systems are given in Section 3.2.

2.2. Utterance Verification (UV)

In **utterance verification**, we are again given an utterance $\mathcal{U}$, represented possibly using a different set and different number of feature vectors $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M\}$, along with a "claimed" (prompted) text, say $k \in \{1, 2, \ldots, K\}$. Now we evaluate

$$\ell_{\text{utt}}(\mathcal{Y}, k) = \log \frac{p(\mathcal{Y}|\text{same text})}{p(\mathcal{Y}|\text{different text})}. \quad (2)$$

In principle, one might use exactly the same methods as in ASV by treating utterances as speakers, i.e. using utterance-specific GMMs adapted from a global UBM. This would induce minimum changes to existing code libraries or be helpful in reducing computation if the same features are shared across ASV and UV subsystems. Alternatively, we may compute the log-likelihood for a given text using forced alignment or an ASR system to compare the decoded output with the claimed text. Details of our UV systems are provided in Section 3.1.

Two key design assumptions are made in the case of our UV systems. The first assumption is *speaker-independence*. The principal motivations for this are user convenience and practicality. In *automatic enrollment* (e.g. [11]) an ASV system interacts with the user without the presence of a human operator, a likely desideratum for any remotely-operated ASV system intended to be transparent or helpful in reducing personnel costs. In such cases, the system should verify the quality (correctness) of the enrolment utterances in terms of their content, even when the system has no prior data from the new speaker, hence excluding a speaker-dependent solution. As a consequence, our UV protocol designed for the RedDots copora (see Section 4.1) contains training speakers that are disjoint from the target and non-target speakers used in ASV evaluation.

Our second assumption is that *the universe of the possible phrases is known (fixed)*, in our case, $K = 10$ common pass-phrases from RedDots part 01. This might correspond, for instance, to the smart-phone unlocking scenario with a publicly known set of phrases, one being randomly prompted to the user at the run-time to prevent a replay attack. Hence, we are allowed to use the other known phrases for score normalization (which we have observed to be crucial). Two normalization techniques are considered. The first, **Mean norm**, subtracts the mean score of the other (competing) phrases from the hypothesized phrase score. The second, **Max norm**, subtracts the maximum.

2.3. Combining ASV and UV systems

In this scenario, given the utterance $\mathcal{U}$, we would verify both its content and the speaker identity and accept the claim only if both are correct, with the goal of tackling replay attacks. The assumption of independent ASV and UV subsystems[1], leads to a product of two likelihood ratios in (1) and (2), thus an additive score $\ell_{\text{utt,spk}}(\mathcal{Y}, \mathcal{X}, j, k) = \ell_{\text{spk}}(\mathcal{X}, j) + \ell_{\text{utt}}(\mathcal{Y}, k)$. Scores may also be weighted as is common for ASV. A key difference from the fusion of heterogeneous ASV classifiers, however, is that here we combine two systems that are designed to solve two different tasks. As a consequence, the usual score fusion approach would provide a possible 'loophole' or hill-climbing attack vulnerability in the case of an attacker who happens to speak the correct pass-phrase, thereby artificially increasing the score and reducing overall security. Hence, we need a fusion strategy that leaves the full system accuracy unaffected in the case of genuine speakers and naive impostors but helps in preventing replay attacks.

[1]This may be a wise assumption given that ASV and UV systems are trained with disjoint speakers and possibly different features and classifiers.

Table 1: *Trial types of RedDots and their categorization for UV, ASV and joint protocol.*

	Speaker Correct	Utterance Correct	True/False Trial for		
			UV	ASV	Joint
Target Correct (TC)	1	1	True	True	True
Target Wrong (TW)	1	0	False	True	False
Impostor Correct (IC)	0	1	True	False	False
Impostor Wrong (IW)	0	0	False	False	False

In this paper, we implemented different types of UV and ASV (as shown in Table 2) system independently developed in three different sites and evaluate their performance for joint verification of spoken content and speaker.

3. System Description

3.1. Utterance Verification Systems

The **UV1** system uses Mel-frequency cepstral coefficients (MFCCs) and a GMM-UBM [12] architecture. MFCCs are extracted using 20 filters in mel scale. After dropping the energy coefficients, we perform RASTA on 19-dimensional coefficients and add deltas and double-deltas to form 57-dimensional features. Finally, we perform utterance-level cepstral mean normalization. For training utterance models, a UBM of 512 components is trained with all the male speech data in TIMIT. The utterance models are obtained using *maximum-a-posteriori* (MAP) adaptation with relevance factor of 3. A target-to-UBM log-likelihood ratio is used as the UV score.

The **UV2** system uses a 2-layer approach based on HMM and a UBM, similar to the 3-layers described in [16] for text-dependent ASV. A left-to-right 5-state HMM is used with continuous observation densities modeled with GMMs adapted from a 512-component UBM trained on TIMIT. To initialise the HMM, one utterance is cut into 5 equal-sized segments, and each HMM's GMM state is estimated by adapting the UBM to its corresponding segment through MAP. Viterbi decoding is then performed to segment the remaining utterances. Finally, the resulting data partitions are used to estimate the final HMM's GMMs by adapting the UBM. The UV score is computed as the likelihood ratio of the phrase-depended HMM aligned by Viterbi decoding and the UBM.

UV3 uses dynamic time warping (DTW) to align feature vectors for a pair of utterances [17]. It uses the same 57 MFCCs as UV1 but without RASTA. Euclidean distance is used as the frame-to-frame distortion. The average score against all the utterance templates is used as the score.

UV4: uses *forced alignment*, a commonly used technique in speech recognition. With the help of acoustic and language models, it searches for the words or phones in a given transcript by aligning the transcribed data with the speech data. We took the 10 sentences and, using TIMIT dictionary, came up with 10 reference transcripts. All the test segments were force aligned with these reference transcripts. The UV score is the average

Table 2: *Summary of the utterance verification and speaker verification systems implemented in this paper.*

Task	Name	Technique	Sequential Information	Feature	Development Data	Transcript
Utterance Verification	UV1	GMM-UBM [12]	No	MFCC	TIMIT	No
	UV2	HMM-UBM [16]	Yes	MFCC	TIMIT	No
	UV3	DTW [17]	Yes	MFCC	-	No
	UV4	Forced Alignment [18]	Yes	MFCC	TIMIT	Yes
Speaker Verification	ASV1	GMM-UBM [12]	No	MFCC	TIMIT	No
	ASV2	GMM-UBM [12]	No	CQCC	TIMIT	No
	ASV3	HMM [19]	Yes	MFCC	RSR2015	No
	ASV4	i-Vector [13]	No	MFCC	RSR2015	No

pseudo-log likelihood of features given the transcript. We use Kaldi [18], based on English language model to match the spoken words or phones.

The acoustic phone model was trained using TIMIT and a standard *deep neural network* (DNN) implementation of Kaldi [18]. The system has a total of 39 English phones in dictionary. MFCCs with linear discriminant analysis (LDA) and feature-space maximum likelihood linear regression (fMLLR) were used as the DNN inputs with left and right contexts of 3 frames. Training consists of three stages: (1) unsupervised training of a stack of restricted Boltzmann machines (RBMs) with 1024 hidden units and 6 hidden layers with 13 training iterations. Next, (2) we train DNN with the objective to classify the individual frames to their correct probability density functions via cross-entropy objective. Finally, (3) we optimize state-level minimum Bayes risk (sMBR) to emphasize state sequences with higher frame accuracy with respect to the reference alignment.

3.2. Speaker Verification Systems

ASV1: This is the same as UV1 except that we adapt target speaker models instead of target phrase models.

The **ASV2** system uses the same back-end as ASV1, but with a different front-end. Here, *constant Q cepstral coefficients* (CQCCs) [20, 21] are used. CQCC is based on the constant Q transform (CQT) [22] widely used in music processing as a variable-resolution time-frequency analysis tool, providing greater frequency and time resolutions at low and high frequencies, respectively. CQCC are obtained by first calculating the CQT power spectrum, followed by a linearisation of the frequency scale followed by discrete cosine transform (DCT) to give 29 static cepstral coefficients. Next, a filter which adaptively emphasizes the articulation rate of the utterance [21] is applied. Then, deltas are computed, resulting in a 58-dimensional features. Non-speech frames are removed by an energy-based speech activity detector. Finally, cepstral mean and variance normalization is applied.

ASV3: A HMM [19] model is trained using speech from many non-target speakers without any speech transcripts. A forced label (e.g., "HI") is assigned to *all training data* during HMM training. The idea is to capture speaker-independent temporal information in the state transition parameters so we call it speaker independent (SI) HMM. Speaker models are derived from the SI-HMM using MAP adaptation [23] of the Gaussian means using the enrollment data. During testing, test data is force-aligned against the target model and the SI-HMM, and log-likelihood ratio is calculated. We found empirically 14 states and 8 mixtures to provide best results.

ASV4 uses i-vectors [13] to represent speech utterances as $\mathbf{S} = \mathbf{m} + \mathbf{T}\mathbf{w}$, where $\mathbf{w}$ is the i-vector, $\mathbf{S}$ is the utterance supervector, $\mathbf{m}$ is the UBM supervector and $\mathbf{T}$ is a low-rank matrix. A gender-dependent GMM-UBM of 512 mixtures with diagonal covariances is trained using 157 male speakers from the RSR2015 corpus consisting of 30 pass phrases from 9 sessions (approximately 42325 utterances) [24]. The i-vector dimension is set at 400, and each target speaker is represented by an average of i-vectors computed over the phrase-wise i-vectors of their enrollment data. Test i-vector is scored against the target speaker specific averaged i-vector (obtained in training phase) using Gaussian probabilistic linear discriminant(G-PLDA)[25]. Before G-PLDA and scoring, i-vectors are length normalized [25]. We re-use the same UBM training data for training total variability space and G-PLDA.

Table 3: *Database description for ASV experiments.*

	Development	Evaluation
Number of Targets	96	152
Target Correct (TC)	1011	1108
Target Wrong (TW)	9099	9972
Impostor Correct (IC)	9059	22220
Impostor Wrong (IW)	81535	200172

Table 4: *Database description for UV experiments.*

	Development	Evaluation
Test Segments	1049	1536
Matched-Text	1049	1536
Unmatched-Text	9441	13824

Table 5: *Utterance verification accuracy (in terms of % EER) for standalone UV experiments.*

System	No Norm		Mean Norm		Max Norm	
	Dev	Eval	Dev	Eval	Dev	Eval
UV1	12.11	9.31	5.72	5.02	2.76	2.08
UV2	5.25	**5.54**	1.74	2.88	0.57	**1.11**
UV3	19.82	24.81	8.19	8.59	6.17	7.80
UV4	19.27	16.60	4.90	5.73	2.76	4.56
Fused	**3.62**	6.13	**1.24**	**2.73**	**0.48**	1.43

4. Experimental setup

4.1. Design of experiments with RedDots Corpora

The experiments are conducted on the speech data available with on-going RedDots challenge[2]. Since the challenge protocol is mainly designed for speaker recognition task, for our experiments, we have prepared protocol with data used in Part 01 of the evaluation containing 10 common phrases [26]. Note that though existing evaluation plan (specially part 04 of the evaluation, i.e., text-prompted condition) can be benefitted with an integrated ASR engine, it does not consider utterance verification as a standalone task, rather it checks whether a spoken-segment matches with the speaker-sentence pair (used for target

[2] https://sites.google.com/site/thereddotsproject/home

Table 6: *Performance on joint protocol (in terms of % of FRR and FAR) on RedDots Part 01 using standalone ASV system. FAR(X) denotes FAR for condition X is the sub-condition of impostor trials.*

System	Dev				Eval			
	FRR	FAR(TW)	FAR(IC)	FAR(IW)	FRR	FAR(TW)	FAR(IC)	FAR(IW)
Standalone ASV:								
ASV1	3.67	18.85	8.59	0.69	1.62	28.70	8.28	0.70
ASV2	3.46	21.18	7.55	0.97	1.53	21.08	**3.60**	0.32
ASV3	2.37	7.68	13.41	0.54	0.63	15.12	10.97	0.78
ASV4	1.98	12.45	8.88	0.06	0.27	16.32	8.78	0.05
ASV Fused	1.29	8.14	5.87	0.01	**0.09**	10.78	4.46	0.01
Standalone UV:								
UV Fused	8.41	**0.00**	91.91	**0.00**	23.74	**0.00**	73.25	**0.00**
Combined UV and ASV:								
Fusion of Fused Scores	**1.29**	5.56	8.54	**0.00**	**0.09**	6.45	6.10	**0.00**
Decision Fusion	8.81	**0.00**	**5.73**	**0.00**	23.74	**0.00**	4.14	**0.00**

enrolment) or not.

Experiments are conducted only on the male speakers due to the lack for female subjects in the corpus. In our protocol, we use ten common phrases from nine different speakers for training utterance models. In total, 1485 sentences are used (roughly 148 files per utterance). The development set is created from the speech-data from 10 speakers where as the evaluation set contains 30 different speakers. The details of the protocol for ASV and UV experiments are described in Table 3 and Table 4, respectively. Since multiple speaker models are created for same phrase, only unique phrase with test segment pairs are used in UV scoring.

4.2. Performance Evaluation

In order to evaluate the performances of UV systems, EER metric is used which represents the error rate when false acceptance probability and false rejection probability are equal. On the other hand, we report FAR and FRR for ASV task. Since the protocol contains three types of impostor trials, we compute three separate FARs, namely FAR(TW), FAR(IC) and FAR(IW), respectively for TW, IC and IW.

5. Experimental Results and Discussion

5.1. Evaluation of Standalone UV System

Reported in Table 5 are standalone UV results for both development and evaluation sets. Among the four systems the HMM-based approach (UV2) gives the lowest error rate. The application of score normalization proves beneficial in all cases with max-based normalization giving better results than mean-based. Results for linear regression based score fusion obtained with the BOSARIS toolkit[3] are also illustrated. Fusion weights are optimized on the development data and applied to the evaluation set without modification. Results for the fused system gives the lowest EER in almost all cases.

5.2. Evaluation of Joint UV-ASV Systems

First, the performance for standalone ASV is reported for the four systems individually and when fused for the joint protocol. These results are shown in Table 6. The FAR for the IW condition is the lowest for most systems. This is expected given the consideration of mismatched-text impostor speakers. The FAR is generally worst for the TW condition which involves the same speakers but wrong utterance (similar to a replay attack). Without the use of UV, this result is also expected. Performance is considerably improved in fused mode, even if improvements are not entirely consistent. Next, performance is reported for standalone, fused UV. While the FAR for TW and IW conditions is 0%, performance is severely degraded for other conditions. This is unsurprising since standalone UV ignores speaker information. Finally, the last two rows of Table 6 report performance for combined ASV and UV. When the two fused systems are combined using score fusion, lower FRRs are obtained, while the FAR for the TW condition remains high. On the other hand, decision-level fusion produces the lowest FAR for the TW condition, albeit with increased FRR.

[3] https://sites.google.com/site/bosaristoolkit/

6. Conclusions

In this paper, we have evaluated joint utterance and speaker verification system on text-dependent RedDots corpora. We have observed considerable improvement in joint verification when UV and ASV systems are used together in a combined mode. This approach could be useful for practical application of voice-based biometrics system, especially in protecting such security systems from playback attack.

7. Acknowledgements

The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

8. References

[1] D. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] J. Joseph P. Campbell, "Speaker recognition: A tutorial," *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437 – 1462, 1997.

[3] M. Przybocki and A. Martin, "NIST speaker recognition chronicles," in *Proc. Odyssey 2004: the Speaker and Language Recognition Workshop*, Toledo, Spain, May 2004, pp. 15–22.

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

[5] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, 2015.

[6] *Speaker Verification Makes Its Debut in Smartphone*. [Online]. Available: http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-02/SpeakerVerificationMakesItsDebutinSmartphone/

[7] 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, 2015.

[8] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Biometrics Special Interest Group (BIOSIG), 2014 International Conference of the*, 2014, pp. 1–6.

[9] Y. Liu, P. Ding, and B. Xu, "Using nonstandard svm for combination of speaker verification and verbal information verification in speaker authentication system," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, 2002, pp. I–673–I–676.

[10] L. Rodriguez-Linares and C. Garcia-Mateo, "A novel technique for the combination of utterance and speaker verification systems in a text-dependent speaker verification task," in *Proc. of ICSLP*, vol. 2, 1998, pp. 213–216.

[11] Q. Li, B.-H. Juang, and C.-H. Lee, "Automatic verbal information verification for user authentication," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 585–596, 2000.

[12] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[13] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011.

[14] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8.

[15] M. Hébert, *Springer Handbook of Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, ch. Text-Dependent Speaker Recognition, pp. 743–762.

[16] A. Larcher, J. Bonastre, and J. S. D. Mason, "Reinforced temporal structure information for embedded utterance-based speaker recognition," in *INTERSPEECH*, Brisbane, Australia, 2008, pp. 371–374.

[17] L. Rabiner and H. Juang B, *Fundamental of speech recognition*. First Indian Reprint: Pearson Education, 2003.

[18] D. Povey, A. Ghoshal, and Others, "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[19] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of the IEEE*, vol. 77, pp. 257–285, 1989.

[20] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop*, Bilbao, Spain, 2016.

[21] ——, "Articulation rate filtering of CQCC features for automatic speaker verification," in *INTERSPEECH*, 2016.

[22] J. Brown, "Calculation of a constant q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, January 1991.

[23] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[24] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent Speaker Verification: Classifiers, Databases and RSR2015," vol. 60, pp. 56–77, 2014.

[25] D. G. Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Proc. of Interspeech*, 2011, pp. 249–252.

[26] K. Lee, A. Larcher, W. Wang, P. Kenny, N. Brummer, D. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Proceedings of Interspeech*, 2015.