Interspeech 2018
2-6 September 2018, Hyderabad

Analysis of the Effect of Speech-Laugh on Speaker Recognition System

Sri Harsha Dumpala, Ashish Panda, Sunil Kumar Kopparapu

TCS Research and Innovation Labs

d.harsha@tcs.com, ashish.panda@tcs.com, sunilkumar.kopparapu@tcs.com

Abstract

A robust speaker recognition system should be able to recognize a speaker despite all the possible variations in speaker's speech. A common variation of the neutral speech is speech-laugh, which occurs when a person is speaking and laughing, simultaneously. In this paper, we show that speech-laugh significantly degrades the performance of an i-vector based speaker recognition system. Further, we show that laughter and neutral speech contain complementary speaker information, which can be combined to improve the performance of the speaker recognition system for speech-laugh scenarios. Using AMI meeting corpus database, we show that by including neutral speech and laughter in enrollment phase, the performance of the system in the speech-laugh scenarios can be relatively improved by 36% in EER.

Index Terms: Laughter, speech-laugh, speaker recognition

1. Introduction

The flexibility of the human speech production system allows production of several variants of neutral speech, such as speech-laugh, depending on the emotional and physical state of the speaker along with non-speech sounds such as laughter, crying, etc. These variations, being produced by the speech production system of the same speaker, are likely to carry certain speaker-specific information. However, it is not immediately clear whether this speaker-specific information is the *same* as the speaker-specific information represented by neutral speech.

Speaker recognition refers to the task of identifying the speaker using speech as the only cue [1]. In recent years, speaker recognition systems have shown significant improvement in performance making them more viable for commercial applications. Current state-of-the-art speaker recognition systems employ i-vector-based approaches, where the feature vectors representing the speech signal are characterized by low-dimensional fixed length vectors called identity-vectors or i-vectors [2], [3], [4]. These i-vectors are obtained by projecting the speech signal onto a subspace $\mathcal{T}$, referred to as "total variability space", which contains both speaker and channel variabilities, simultaneously [2]. Approaches based on i-vectors are well established as they have shown a significant improvement in speaker recognition performances. However, the effect of the variations of the neutral speech on i-vector-based systems have not been reported. More explicitly, there is a need to identify if the *neutral* speech of a speaker is sufficient to capture the speaker-specific characteristics completely, irrespective of the variations in the speech produced by the speaker in natural day to day conversations. This information is essential for developing robust speaker recognition systems.

Only a few studies have analyzed the effect of different non-speech sounds such as breath, whistle, etc., on the performance of speaker recognition systems [5], [6], [7]. It is evident from earlier studies that the performance of the speaker recognition system trained using only neutral speech of the speaker degrades, if these non-speech sounds are a part of the testing phase of the system [6], [7]. Most of these studies considered traditional Gaussian mixture models with universal background model (GMM-UBM) for their study.

In this paper, we study the effect of speech-laugh on the state-of-the-art i-vector speaker recognition system. While earlier studies have focused on only the non-speech sounds (such as breaths, whistle etc.), this study focuses on the non-speech event co-occurring with speech (i.e. speech-laugh: laugh co-occurring with speech). This is important, since in natural conversations, a significant part of the non-speech sounds co-occur with speech to produce variants of neutral speech such as speech-laugh, breathy speech, etc., [8], [9], [10]. The main objectives of this analysis are twofold:

1. To investigate whether the i-vectors extracted from the neutral speech of a speaker are robust to the variations in speech produced by the speaker.
2. To analyze the importance of including non-speech sounds, such as laugh, along with neutral speech in the enrollment phase of speaker recognition, when variants of neutral speech are present in the testing phase.

To achieve these objectives, laughter (a non-speech sound) and speech-laugh (a variant of neutral speech) segments produced by the speaker are considered to evaluate the performance of speaker recognition system developed using only neutral speech. Further, the performance of speaker recognition system, particularly when tested with speech-laugh, is analyzed when laughter sounds collected from the speaker are included in the enrollment phase. These systems are evaluated on neutral speech, laughter and speech-laugh of the speaker to ascertain the presence of any complementary speaker-specific information provided by laughter. Studying the effect of laughter and speech-laugh is important, as laughter is one of the most common non-speech sound which occurs very frequently [11], and in natural conversations, more than 50% of these laughter sounds happen to be speech-laughs [12].

The organization of the paper is as follows. The approach followed for analysis is explained in Section 2. Section 3 summarizes the dataset and the i-vector-based speaker recognition system considered. Experimental results are given in Section 4. Summary along with conclusions are given in Section 5.

2. Background and Proposed Approach

Analysis is performed by considering neutral speech (NS), laughter (L) and speech-laugh (SL) produced by the speakers.
Neutral speech (NS) refers to the normal/regular speech of the speaker, which carries mostly linguistic information.
Laughter (L) is a non-speech sound, typically produced by a series of sudden bursts of air through the vocal tract system [13], [14], [15]. Laughter does not carry any linguistic information but might provide important speaker-specific cues.
Speech-laugh (SL) refers to the segments of speech, where

10.21437/Interspeech.2018-2090

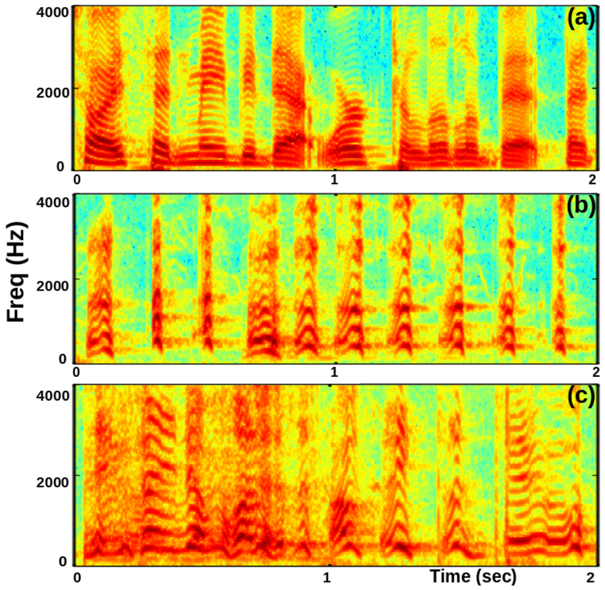

Figure 1: *Spectrograms obtained for (a) neutral speech, (b) laughter and (c) speech-laugh, respectively.*

laughter co-occurs with neutral speech [16], [17]. Speech-laugh exhibits characteristics of both laughter and neutral speech, but it is very distinct from both, laughter and neutral speech [12], [17], [18]. Hence, speech-laugh forms a separate class, which carries both, linguistic and non-linguistic information [18].

The spectral features obtained from neutral speech (NS), laughter (L) and speech-laugh (SL) form a continuum, with laughter exhibiting higher formant frequencies (particularly, first formant frequency) followed by speech-laugh and then neutral speech [17], [19]. This can be observed from the spectrograms obtained for neutral speech, laughter and speech-laugh samples of the same speaker as shown in Fig. 1. This variation in formant frequencies, which might carry speaker-specific information [20], can affect the performance of the speaker recognition systems trained on only neutral speech but tested on more natural speech which also consists speech-laugh and laughter segments. The effect of such variations on speaker recognition system is analyzed in this study.

The approach followed for the analysis is depicted in Fig. 2. It can be observed from Fig. 2 that apart from neutral speech, laughter data collected from the speakers is also included in the enrollment phase of the speaker recognition system. For analysis, four different i-vector-based systems are developed: first system using only neutral speech (NS), second system using only laughter (L) and the third one is developed considering both, neutral speech and laughter (NS $\cup$ L) of the speakers as enrollment data. Further, a fourth system obtained by late fusion (of the output scores of the first and the second systems) is also considered to analyze the variation in speaker-specific information provided by laughter and neutral speech. The effect of including laughter in the enrollment phase is analyzed by evaluating the performance of these systems on speech-laugh (which is not used in the enrollment phase of any system) of the speakers. This analysis is significant for two main reasons. First, it shows that a variant of neutral speech (i.e. speech-laugh) can be handled by simply including a non-speech event (i.e. laugh) in the enrollment phase. Second, laugh may be easier to generate during enrollment than speech-laugh as subjects may not speak while laughing to produce speech-laugh.

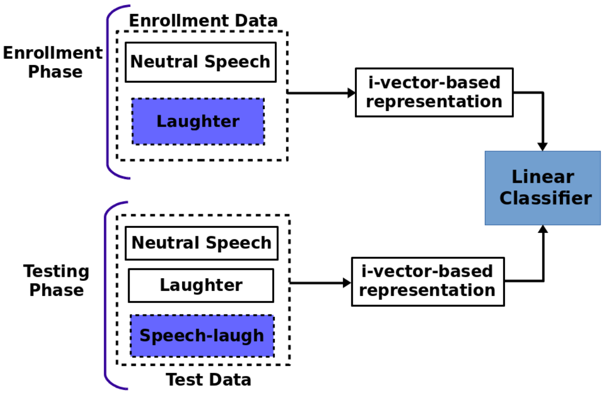

Figure 2: *Block diagram of the approach followed for analysis.*

3. Experimental details

3.1. Database details

For this analysis, GMM-UBM and i-vector statistics as provided in the Voice biometry standardization (VBS) [21] toolkit are used. In VBS toolkit, GMM-UBM with 2048 components was trained using NIST SRE 2004 − 2008 data ($\approx$ 1156.03 hours of data) and the $\mathcal{T}$ matrix required for i-vector extraction was trained using Fisher English (Part 1 and 2), NIST SRE 2004 − 2008, and Switchboard corpus (Phase 2, Phase 3, cellular part 1 and cellular part 2) which totals to 9010.23 hours of data. But standard speaker recognition corpus such as NIST SRE does not include speech transcripts, especially for non-speech sounds. Hence, the enrollment and the test i-vectors used in this analysis are obtained using AMI meeting corpus [22].

AMI meeting corpus is a multi-modal dataset consisting of 100 hours of meeting recordings. Each meeting includes four speakers discussing spontaneously on a given topic in English. Most of the speakers are non-native English speakers, thus providing a high degree of variability in speech. Each speaker's audio was recorded using individual headset condenser microphones and lapel microphones. We used the data collected through headset condenser microphones as speech of the considered speaker is profound. The corpus is labeled at word level. Further, the laughter segments produced by the speaker are separately labeled along with their timestamps. For this analysis, we considered 100 speakers (60 female and 40 male) whose recordings have significant amount of laughter and speech-laugh content.

3.2. Data Organization

For the purpose of analysis, we organized the audio corpus into 4 different datasets (see Table 1). The speech of the speaker, in the corpus, is marked as neutral speech (NS), laughter (L) and speech-laugh (SL). It can be observed from Table 1 that enrollment sets (namely, ES1, ES2 and ES3) of DSET1 (NS), DSET2 (L) and DSET3 (NS $\cup$ L) are used in enrollment phase of the speaker recognition system, whereas test sets of all datasets (i.e., TS1 through TS4) are used in the testing phase. The enrollment set in DSET3 (namely, ES3) consists a total of 85 utterances (70 NS utterances and 15 L utterances) each of 3.5 sec to 4 sec in duration. But every utterance in TS3 contains both NS and L, where the proportion of laughter (L) typically varies from 20% to 50%. It is to be noted that SL is used only in the testing phase (TS4), but not in the enrollment phase.

Table 1: *Dataset organization details (Enrollment (ES) and Test (TS) refer to Enrollment set and Test set, respectively).*

Dataset	Contents	# Utterances/speaker		Duration (sec)
		Enrollment (ES)	Test (TS)	
DSET1	Neutral speech (NS)	ES1 = 80	TS1 = 25	3.5-4
DSET2	Laughter (L)	ES2 = 20	TS2 = 10	3.5-4
DSET3	Neutral speech and laughter (NS ∪ L)	ES3 = 85	TS3 = 10	3.5-4
DSET4	Speech-laugh (SL)	–	TS4 = 10	3.5-4

3.3. System description

The i-vector-based speaker recognition systems considered for analysis are implemented using the VBS [21] toolkit. Figure 3 shows the schematic of the system implementation using VBS and consists of the audio, voice activity detection (VAD), feature extraction, i-vector extraction and post processing. We use probabilistic linear discriminant analysis (PLDA) as the metric to measure the performance of the speaker recognition system. We briefly describe the blocks in VBS system [21] as used in our experiments.

Audio: The audio data considered in the enrollment phase (see Enroll Audio in Fig. 3) consists of NS and L sounds of each speaker. Whereas the audio data in the testing phase (see Test Audio in Fig. 3) consists of NS, L and SL. All the audio samples are down sampled to 8 kHz and are in 16-bit PCM format as required by the VBS.

Voice activity detection (VAD): VAD is used prior to feature extraction to remove the silence and low signal-to-noise ratio (SNR) regions in the audio sample. In this analysis, VAD is performed using the VOICEBOX toolkit [23]. It is to be noted that most of the speech-laugh segments are voiced [9] and cannot be eliminated by using a conventional VAD.

Feature extraction: The audio signals are represented using mel-frequency cepstral coefficients (MFCCs), which are widely used in i-vector-based speaker recognition systems [2], [3]. MFCCs are extracted using 25 msec Hamming window with 10 msec forward shift. MFCCs are computed by using 24 mel-filter banks and limiting the bandwidth to frequency in the range of 125 Hz - 3800 Hz. Every frame is represented using 20 coefficients (first 19 MFCCs along with the 0^{th} coefficient). This 20-dimensional feature vector is mean and variance normalized using a 1 sec sliding window. Subsequently, the delta and the double delta coefficients are computed to form a 60-dimensional feature vector to represent each frame.

i-vector extraction: To obtain a low-dimensional fixed-length i-vector-based representation of the sequence of feature vectors, the GMM-UBM and the i-vector statistics (total variability space ($\mathcal{T}$)) are necessary. In this analysis, the GMM-UBM-based i-vectors are extracted using the GMM-UBM and $\mathcal{T}$ matrix statistics released by VBS. The gender-independent universal background model (UBM) with 2048 components and the total variability space $\mathcal{T}$ of 600-dimension were trained using the data as explained in Section 3.1.

i-vector post-processing: The i-vectors of 600-dimensions obtained for each audio sample are reduced to 200-dimensions using linear discriminant analysis (LDA) [2], [24]. Then these i-vectors are further normalized using within-class covariance matrix [24]. Both, LDA and within-class covariance matrix are provided by VBS and are trained on the same data that is used for $\mathcal{T}$ matrix enrollment. In this analysis, the speaker templates (namely, i-vectors corresponding to each speaker) are generated separately for the three considered cases (namely, NS, L and NS ∪ L) during the enrollment phase, and an i-vector is obtained for each audio sample during the test phase.

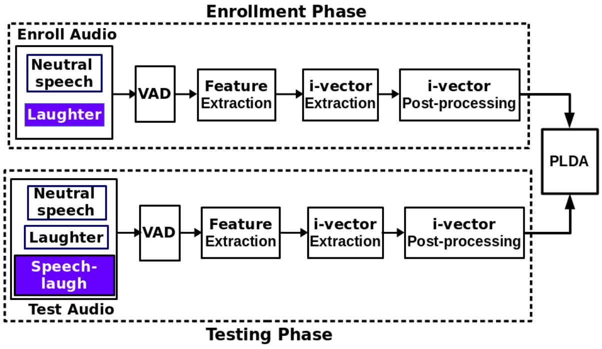

Figure 3: *Block diagram of i-vector-based speaker recognition system implementation.*

Table 2: *Details of the considered speaker recognition systems.*

System	Training set	Test sets
SYSTEM1	ES1 (NS)	TS1, TS2, TS3, TS4
SYSTEM2	ES2 (L)	TS1, TS2, TS3, TS4
SYSTEM3	ES3 (NS ∪ L)	TS1, TS2, TS3, TS4
SYSTEM4	Fusion of SYSTEM1 and SYSTEM2	TS1, TS2, TS3, TS4

PLDA: To compare the enrollment i-vectors to the test i-vectors for speaker recognition, PLDA is used [25], [26]. PLDA is a special case of joint factor analysis (JFA) with single Gaussian component, but is used in the i-vector space. Given a pair of i-vectors, PLDA computes the log-likelihood score for the same-speaker and the different-speaker hypothesis [27]. This PLDA score is used to evaluate the speaker recognition systems.

4. Experimental results

The performance of the i-vector-based speaker recognition systems (see Table 2) is evaluated in terms of Equal Error Rate (EER); lower EER value indicates better performance of the system. The EER (in %) obtained for the four considered systems, namely, SYSTEM1 (trained on ES1 i.e., neutral speech only), SYSTEM2 (trained on ES2 i.e., laughter only), SYSTEM3 (trained on ES3 i.e., neutral speech and laughter), along with SYSTEM4 (fusion of SYSTEM1 and SYSTEM2) when tested on all the test datasets (TS1 through TS4) are shown in Table 3 (refer to Table 1 for dataset details).

SYSTEM4 is obtained by late fusion of the PLDA scores obtained by SYSTEM1 and SYSTEM2. The fusion of the PLDA scores is performed as shown in Eq. (1),

$$PLDA_{S4} = (\alpha \times PLDA_{S1}) + ((1-\alpha) \times PLDA_{S2}), \quad (1)$$

where $PLDA_{S1}$ and $PLDA_{S2}$ represents the PLDA scores of SYSTEM1 and SYSTEM2, respectively and α is a weight (which gives higher importance to one system and lower importance to other system. $\alpha = 0.5$ means both systems contributes equally for the final output) whose value ranges from $0-1$ (see Table 4). EER for System4 is computed by using the $PLDA_{S4}$ scores obtained on the test set. The overall best performance (in EER) obtained by System4 (i.e., at $\alpha = 0.5$) is provided in Table 3. As can be observed from Table 3

- The matched scenarios (denoted by (*) in Table 3) always work better than the mismatched condition, which is along the expected line.
- When tested on only neutral speech (Ts1), System1, System3 and System4 performs better than System2, which is trained only on laughter sounds. Similarly, when tested on utterances with only laughter (Ts2), System2, System3 and System4 performs better than System1. This shows that the i-vector-based speaker representation obtained from the neutral speech (Ns) of a speaker differs from that of laughter, signifying the variation in speaker-specific information exhibited by neutral speech and laughter.
- When tested on speech-laugh (Ts4), the performance of System1 (EER = 16.00%) degrades compared to its performance on test sets with neutral speech i.e., Ts1 and Ts3. Also, the performance of System2 (EER = 15.61) degrades when tested on speech-laugh, compared to its performance on test sets with laughter i.e., Ts2. But, the performance of both System1 and System2 is very close on Ts4. This shows that speech-laugh exhibits characteristics of both, laughter and neutral speech, but also is distinct from both laughter and neutral speech. Better performance of System2 compared to System1 on Ts4 may be attributed to the small laughter segments occurring within speech-laugh segments.
- System3 performs better than System1 and System2 on test sets Ts1 and Ts3 . Further, when tested on speech-laugh (Ts4, which is not present in any enrollment set), System3 (EER = 11.73%) performs better than System1 (EER = 16.00%) and System2 (EER = 15.61%). Better performance of System3 compared to System1 on Ts1 might be attributed to the additional complementary information provided by laughter which better handles the variations in neutral speech.
- System4 which is fusion of System1 and System2, outperforms all other systems on the test sets, Ts3 and Ts4. In particular, when tested on speech-laugh (Ts4), System4 (EER = 10.16%) outperforms System1 by 5.84%, System2 by 5.45% and System3 by 1.57% in EER. This shows that the speaker-specific information captured by System1 and System2 when fused may be able to better represent the speaker-specific information embedded in speech-laugh.
- When tested on laughter (Ts2), System4 (EER = 11.82) outperforms System3 by 7.66%. This may be attributed to the higher proportion of neutral speech in train set Es3 compared to laughter, whereas for System4, equal weightage is given to both neutral speech (System1) and laughter (System2).

The better performance of System4 and System3 when tested on Ts4 (i.e., speech-laugh) compared to System1 and System2, shows that laughter and neutral speech might carry complementary speaker-specific information which when combined, helps in improved speaker recognition performance on speech-laugh.

Table 3: *EER (in %) obtained for the systems on different datasets.* Ns, L *and* Sl *refers to Neutral speech, laughter and speech-laugh, respectively, and* Syst *refers to* System. *(*) denotes matched train and test conditions.*

Test-set	Syst1 (Es1)	Syst2 (Es2)	Syst3 (Es3)	Syst4 (Fusion)
Ts1 (Ns)	2.95*	23.78	**2.93**	4.18
Ts2 (L)	29.37	**9.17***	19.48	11.82
Ts3 (Ns ∪ L)	12.93	18.09	7.15*	**6.77**
Ts4 (Sl)	16.00	15.61	11.73	**10.16**

Table 4: *EER (in %) obtained for* System4 *for different α*

α	Ts1	Ts2	Ts3	Ts4
0	23.78	**9.17**	18.09	15.61
0.1	18.09	10.35	17.94	13.81
0.2	12.54	10.87	13.85	12.26
0.3	9.04	11.09	10.47	11.61
0.4	5.92	11.63	9.10	10.91
0.5	4.18	11.82	6.77	**10.16**
0.6	3.49	15.60	**5.61**	10.28
0.7	3.10	19.43	6.79	10.74
0.8	**2.81**	22.70	8.36	12.00
0.9	2.84	26.65	10.20	13.78
1	2.95	29.37	12.93	16.00

Table 4 shows the performance of System4 by varying α in Eq. (1) from $0-1$. As α move towards 1, the performance of System4 improves on neutral speech (Ts1), and degrades on laughter (Ts2), as expected. But an improved performance on Ts3, Ts4, and an improvement in overall system performance is observed when α is close to 0.5 (i.e., $\alpha = 0.4, 0.5, 0.6$). This shows that laughter and neutral speech carry complementary speaker-specific information, and also signifies the importance of this speaker-specific information for improved speaker recognition performance on speech-laugh.

5. Conclusions

Natural conversations between people have significant amount of non-speech events, such as laughter, which co-occur with speech. A practical speaker recognition system needs to be able to recognize a speaker in these scenarios. In this paper, we show that the i-vector-based speaker recognition system trained on neutral speech performs poorly when it encounters speech-laugh in the test utterances. We also show that laughter and neutral speech contain complementary speaker-specific information, which can be combined to deal with speech-laugh. Towards this end, we propose the use of laughter along with neutral speech in the enrollment data. This can be accomplished in two ways. One, by simply pooling laughter and neutral speech data to train the enrollment i-vector or two, by training separate i-vectors for neutral speech and laughter and fusing their PLDA scores. Currently, we are extending this study to other non-speech events. We are also studying better ways of score-fusion, which can improve the performance of the system.

6. References

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

[3] O. Glembek, P. Matějka, O. Plchot, J. Pešán, L. Burget, and P. Schwarz, "Migrating i-vectors between speaker recognition systems using regression neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] A. K. Sarkar, D. Matrouf, P. M. Bousquet, and J. F. Bonastre, "Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification." in *Interspeech*, 2012, pp. 2662–2665.

[5] A. Janicki, "On the impact of non-speech sounds on speaker recognition," in *International Conference on Text, Speech and Dialogue*. Springer, 2012, pp. 566–572.

[6] S. H. Dumpala and R. K. Alluri, "An algorithm for detection of breath sounds in spontaneous speech with application to speaker recognition," in *International Conference on Speech and Computer*. Springer, 2017, pp. 98–108.

[7] M. K. Nandwana, H. Boril, and J. H. Hansen, "A new front-end for classification of non-speech sounds: a study on human whistle." in *INTERSPEECH*, 2015, pp. 1982–1986.

[8] S. H. Dumpala and S. K. Kopparapu, "Improved speaker recognition system for stressed speech using deep neural networks," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 1257–1264.

[9] S. H. Dumpala, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Use of vowels in discriminating speech-laugh from laughter and neutral speech," *Interspeech 2016*, pp. 1437–1441, 2016.

[10] H. Hirose, "Investigating the physiology of laryngeal structures," *The handbook of phonetic sciences*, pp. 130–52, 2010.

[11] K. P. Truong and D. A. Van Leeuwen, "Automatic detection of laughter." in *INTERSPEECH*, 2005, pp. 485–488.

[12] E. E. Nwokah, H. Hsu, P. Davies, and A. Fogel, "The integration of laughter and speech in vocal communicationa dynamic systems perspective," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 4, pp. 880–894, 1999.

[13] C. Wallace, "The phonetics of laughter–a linguistic approach," in *Interdisciplinary Workshop on the Phonetics of Laughter*, 2007, pp. 4–5.

[14] A. Batliner, S. Steidl, F. Eyben, and B. Schuller, "On laughter and speech laugh, based on observations of child-robot interaction," *The phonetics of laughing, trends in linguistics. de Gruyter, Berlin, to appear*, 2010.

[15] J. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *The Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1581–1597, 2001.

[16] J. Trouvain, "Phonetic aspects of speech-laughs," in *Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: LHarmattan*, 2001, pp. 634–639.

[17] C. Menezes and Y. Igarashi, "The speech laugh spectrum," *Proc. Speech Production, Brazil*, pp. 157–164, 2006.

[18] S. H. Dumpala, K. V. Sridaran, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of laughter and speech-laugh signals using excitation source information," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 975–979.

[19] D. P. Szameitat, C. J. Darwin, A. J. Szameitat, D. Wildgruber, and K. Alter, "Formant characteristics of human laughter," *Journal of voice*, vol. 25, no. 1, pp. 32–37, 2011.

[20] J. Franco-Pedroso and J. Gonzalez-Rodriguez, "Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors," *Odyssey 2016*, pp. 237–244, 2016.

[21] "Voice biometry standardization (VBS) initiative," http://voicebiometry.org/, 2015.

[22] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[23] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, vol. 47, 1997.

[24] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, vol. 2011, 2011, pp. 249–252.

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

[26] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[27] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matějka, and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4832–4835.