An investigation of acoustic features of the lower vocal tract for speaker recognition

Peter Birkholz, Xinyu Zhang

Institute of Acoustics and Speech Communication, TU Dresden
peter.birkholz@tu-dresden.de

Abstract: Speaker recognition systems often use mel-scaled cepstral coefficients (MFCCs) as main features. In contrast to MFCCs, Godoy et al. (2015) proposed a different type of short-term spectral analysis that provides features related to the lower vocal tract (LVT). They are calculated as the ratio of the acoustic short-time spectra during the closed and open phases of the glottal oscillation cycles based on a pitch-synchronous analysis. These features were suggested to be particularly speaker-specific and might therefore be suitable to substitute or complement MFCCs in speaker recognition systems. The present study investigated the benefit of these features in an i-vector-based speaker recognition system. Using the LVT features alone, the system achieved a speaker recognition rate of 92.3% with 63 enrolled speakers. When the LVT features were fused with conventional MFCC features, the recognition rate was about equal to the recognition rate using MFCC features alone ($> 98\%$).

1 Introduction

Differences in vocal tract dimensions, vocal fold characteristics, and other parts of the speech production system give each individual a unique voice. Speaker recognition systems attempt to extract these speaker-specific features from the speech signal to identify the speaker [1]. Most speaker recognition systems use short-term spectral features like mel-frequency cepstral coefficients (MFCCs) [2], perceptual linear prediction coefficients [3], or linear predictive cepstral coefficients [4]. Some systems also use prosodic features [5, 6] or voice source features [7, 8] to supplement the short-term spectral features and to improve the recognition accuracy.

In addition to these commonly used features, there are other characteristics that might be useful for speaker recognition. Titze [9] discovered in 1996 that the properties of the lower vocal tract (epilaryngeal tube and piriform fossae) affect the formant frequencies of the vocal tract in specific ways. Also Honda et al. [10] found that the resonances of the hypopharyngeal cavity significantly impact the spectrum in the high frequency region. Since the shape of the hypopharyngeal cavity is more or less unique for a specific speaker and relatively stable across different phonemes, these characteristics might be useful features for speaker recognition. In 2015, Godoy et al. [11] proposed a type of short-term spectral analysis that reveals acoustic features related to the lower vocal tract (LVT). The analysis determines the spectral difference of the audio signal during the open and closed phases of the glottal oscillation cycle. This closed-open spectral difference highlights the resonances of the epilaryngeal tube and the piriform fossae, which are related to the speaker-specific sizes of these cavities [12].

In the present study we examined the benefit of these LVT features for text-independent speaker recognition, both alone and in combination with conventional MFCC and delta-MFCC features. As the classifier, we used the well-established i-vector model [13].

DOI: 10.35096/othr/pub-7087, URN: urn:nbn:de:bvb:898-opus4-70870

2 Method

2.1 Corpora

The method to calculate the LVT features described below requires the determination of the glottal closure instants (GCIs). There are multiple algorithms to estimate GCIs directly from the audio signal [14], but they can be determined most accurately by means of the electroglottography (EGG) signal [15]. To prevent potential errors of an audio-based GCI estimation to bias the results of this study, we therefore decided to determine the GCIs from EGG signals. Hence, we used speech corpora that contained both the audio and the EGG signal.

2.1.1 Winkler corpus

The Winkler corpus [16] was collected by Ralf Winkler for his dissertation and contains speech material of 68 German speakers between 18 and 82 years old. The EGG and the audio signals were digitally and synchronously recorded at a sampling rate of 44100 Hz and with 16-bit quantization in a quiet room. The audio signal was recorded with a high-quality headset (AKG C 410), and the EGG signal with the Laryngograph (Laryngograph Ltd).

The material of each speaker contains sustained speech sounds, different logatomes, a read news text, and spontaneous speech (a description of a picture or of the previous weekend). For the present study, the recordings of the read text (4 long sentences, spoken in 27–50 s) and the spontaneous speech (with durations from 33–192 s) were used and manually split into utterances (pseudo-sentences) of 3–9 s each. The splitting was performed at speech pauses that do not necessarily coincide with the beginnings or endings of linguistic sentences. The material of 9 speakers was omitted due to a low quality of the EGG signal or incomplete data. From the remaining 59 speakers (26 male, 33 female), a total of 964 utterances were obtained. For each speaker, 8 utterances were used for speaker enrollment and testing, with almost equal proportions of read and spontaneous speech. The remaining 492 utterances were chosen for training the universal background model (UBM, see below) of the speaker recognition system.

2.1.2 BITS Unit Selection corpus

The BITS Unit Selection corpus (BITS-US) contains EGG and audio recordings of read speech of two female and two male German speakers [17] with 1683 sentences per speaker. The audio signals were recorded with a close talk microphone (Beyerdynamic NEM 192) and the EGG signal was recorded with the LaryngoGraph PCLX, both at a sampling rate of 48 kHz and with 16 bit quantization. Sixteen randomly selected sentences from each speaker were used in the present study: 8 sentences for the training of the UBM, and 8 sentences used for speaker enrollment and testing. The durations of these partitions were similar to that of the Winkler corpus.

In summary, the material used from both corpora comprised utterances of a total of 63 speakers, with an average of 100 s of speech per speaker.

2.2 Glottal closure instant detection

The glottal closure instant detection and the feature extraction based on this were implemented with a custom-made script for Matlab R2021a. For all used utterances of both corpora, the glottal closure instants were determined as the peaks of the first derivative of the corresponding EGG signals [15]. To this end, the raw EGG signal of each utterance was first downsampled to the common sampling rate of 16 kHz using the function `resample` of Matlab. Then, low-frequency noise was removed from the signal with a zero-phase finite-impulse-response high-pass filter with a stopband from 0 Hz to 50 Hz, and a passband from 60 Hz to 8 kHz. The first derivative of the filtered signal was approximated by finite differences. In the resulting differentiated EGG signal (dEGG), all local maxima above a threshold of 20% of the maximum positive value of the whole signal were considered as GCIs [18, 19]. The delay between the EGG signals and the audio signals due to the sound propagation time from the glottis to the microphone was compensated for in all recordings. As an example, Figures 1 (a) to (c) show the high-pass-filtered EGG signal, the dEGG signal, and the time-aligned speech signal of 30 ms of voiced speech, where the dash-dotted lines are the detected GCIs.

2.3 Feature extraction

Like the EGG signals, the speech signals of both corpora were first resampled to the common sampling rate of 16 kHz. Then the MFCC and LVT features were extracted from the signals using overlapping frames with a frame length of 30 ms and a step size of 10 ms. Only voiced frames that contained at least three GCIs (i.e. two complete glottal periods) were used for feature extraction. Although MFCCs can also be obtained for voiceless frames, this is not possible for LVT features (see below). Hence, we omitted all frames for which we could not determine the complete set of LVT and MFCC features.

For each valid frame, 20-dimensional MFCCs were extracted with the Matlab function `mfcc` using a 512-point FFT, where the first coefficient was replaced by the log energy. In addition, the delta MFCCs were calculated as the central difference between the MFCCs of the following and previous frames.

The calculation of the LVT features is illustrated in Figure 1 (d–h). First, the fundamental period between the two GCIs around the middle of the frame was selected for the further analysis. This period of length T_0 was split into two parts with the lengths $CQ \cdot T_0$ for the closed phase of the glottis, and $(1 - CQ) \cdot T_0$ for the open phase of the glottis, where CQ is the assumed glottal closed quotient, which is a hyperparameter in this study. Then, two Hamming windows were applied to the closed and open phase segments, respectively. The windowed signals were zero-padded to the length of 256 points to calculate the 128-point magnitude spectra of the closed and open phases using the Fast Fourier Transform. The difference spectrum was then calculated as the difference between the two spectra with *logarithmic* amplitudes (in dB). Finally, the Discrete Cosine Transform was applied to the difference spectrum to de-correlate the data. The first 20 DCT coefficients were taken as the LVT feature vector. Finally, the MFCCs and LVTFs were mean- and variance-normalized over a sliding window of three seconds.

2.4 Speaker recognition system

The well-established i-vector model [13] was used to investigate the closed-set speaker recognition performance for different combinations of the feature sets described above. In the i-vector framework, the utterances of a speaker (and hence the speaker itself) are modeled as the probability density function of their feature vectors using a Gaussian Mixture Model (GMM) with a fixed number of Gaussians. The means of the Gaussians are concatenated to form a supervector **s**, which is in turn modeled as

$$\mathbf{s} = \mathbf{m} + T\mathbf{w}, \tag{1}$$

where **m** is the speaker- and channel-independent supervector (i.e., the UBM supervector [20]), T is a matrix of bases spanning the subspace covering the speaker- and session-variability in the supervector space, and **w** is a standard normally distributed latent variable. For each feature

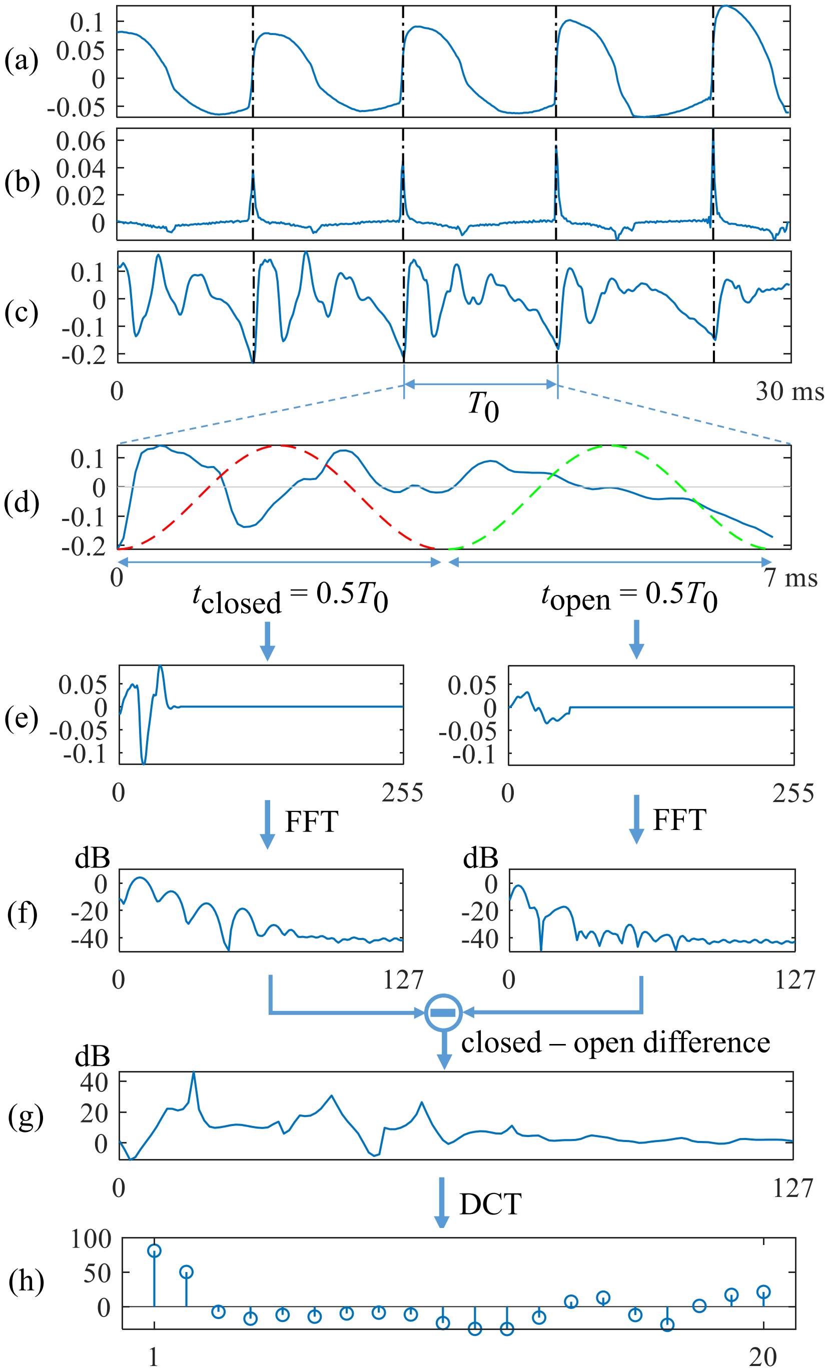

Figure 1 – An illustration of LVT feature extraction for a frame with 30 ms of voiced speech. (a) EGG signal. (b) Differentiated EGG signal. (c) Speech signal (GCIs are marked with vertical lines). (d) Speech signal of the middle glottal period with windows (red and green) for the closed and open phases for $CQ = 0.5$. (e) Windowed and zero-padded signals of closed and open phases. (f) Log. magnitude spectra of closed and open phases. (g) Closed-open spectral difference. (h) 20-dimensional DCT coefficients.

vector sequence representing an utterance or speaker, the i-vector is a Maximum A Posteriori (MAP) point estimate of **w**. The subspace of the i-vectors is usually much lower-dimensional than the supervector space. Finally, a linear discriminant analysis (LDA) and within-class covariance normalization is used to compensate for the channel-dependent information in the i-vector space. The dimension of the LDA is again lower than that of the i-vectors. To determine the most-likely speaker of a given utterance, the cosine similarities of the i-vector of the utterance and the i-vectors of the enrolled speakers (all projected into the LDA space) are calculated, and the speaker with the highest similarity score is selected.

Table 1 – Recognition accuracy (ACC) of i-vector systems with different features and configurations.

Features	# Gaussians	i-vector Dim	LDA Dim	ACC
20 dim. LVTF	128	80	40	92.3%
20 dim. MFCC	64	80	40	97.0%
20 dim. MFCC + 20 dim. LVTF	64	80	40	98.6%
20 dim. MFCC + 20 dim. ΔMFCC	64	80	40	98.6%
20 dim. MFCC + 20 dim. ΔMFCC + 20 dim. LVTF	64	100	50	98.8%

2.5 Experiments

The system described above was used with five different sets of features, which are given in the left column of Table 1. For each feature set, a full grid search was used to tune the hyperparameters of the recognition system. The hyperparameters and their tested values include the closed quotient with $CQ \in \{0.45, 0.5, 0.55\}$, the number of Gaussian components in the GMMs $\in \{64, 128, 256\}$, and the dimensions of the i-vectors $\in \{60, 80, 100\}$. The dimension of the LDA space was always set to half the dimension of the i-vector space. The system performance for each condition was evaluated using 4-fold cross-validation. In each iteration of the cross-validation, 6 out of 8 utterances per speaker were used for enrollment (to determine the i-vector of the speaker), and the remaining 2 utterances were used for testing.

3 Results and discussion

The experimental results are shown in Table 1. For each set of features, the 2nd to 4th columns show the hyperparameter values that gave the best recognition rate shown in the rightmost column. The optimal value for CQ was 0.55 in all cases. Using LVT features alone, an accuracy of 92.3% was achieved. This suggests that these features contain essential information that allows speakers to be distinguished from one another. However, using MFCC features alone, a recognition rate of 97.0% was achieved, which is 4.7% higher than that with LVT features. The recognition accuracy with both LVT and MFCC features combined was 98.6%, slightly higher than with only the MFCC features. However, this increase is too small to claim that adding LVT features to the "conventional" MFCC features generally improves the speaker recognition performance. Instead, we have a ceiling effect here, where the accuracy based on MFCC features alone is already so high that a potential benefit of the LVT features cannot be clearly determined. The same holds when we include ΔMFCC features, as shown in the last two rows of Table 1.

Finally, we compare the results obtained with the LVT features with some other studies using "alternative" features for speaker recognition. For example, Jawarkar et al. [21] proposed gammatone-frequency cepstral coefficients (GFCC) as features and achieved a recognition rate of 93.91% on a corpus with 80 speakers. Daqrouq and Tutunji [22] proposed a method for speaker feature extraction combining formants, wavelet entropy and neural networks. They obtained a recognition rate of 90.09% on a corpus of 80 speakers using an only 12-dimensional feature vector. More recently, Nassif et al. [23] proposed a speaker identification algorithm based on computational auditory scene analysis (CASA) and a cascaded GMM-CNN classifier.

In the neutral talking condition, they achieved a recognition rate of 95.4% on a dataset of 32 speakers. Our recognition rate of 92.3% with only the LVT features is quite comparable to these other studies.

4 Conclusions

This study has shown that the pitch-synchronous spectral analysis proposed by Godoy et al. [11] does provide speaker-specific information that is useful for speaker recognition. However, we could not demonstrate that these new features increase the speaker recognition accuracy when used in combination with conventional MFCC features, as performance based on MFCC features alone was already near prefect. To show the potential benefit of including LVT features in the future, the recognition performance of the MFCC baseline system would have to be lower. This could be achieved by increasing the number of speakers to distinguish, or by increasing the level of difficulty, e.g. by adding noise to the test samples. Finally, to use LVT features in a real-life speaker recognition system, the glottal closure instants would need to be determined from the audio signal without the need for the EGG signal.

5 Acknowledgements

This project was funded by the Federal Ministry for Economic Affairs and Energy (BMWi) through the AiF Projekt GmbH under grant no. ZF4443005HB9. We thank Ralf Winkler for providing his corpus of speech audio and EGG signals.

References

[1] Hansen, J. H. and T. Hasan: *Speaker recognition by machines and humans: A tutorial review*. *IEEE Signal Processing Magazine*, 32(6), pp. 74–99, 2015.

[2] Davis, S. and P. Mermelstein: *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357–366, 1980.

[3] Hermansky, H.: *Perceptual linear predictive (PLP) analysis of speech*. *The Journal of the Acoustical Society of America*, 87(4), pp. 1738–1752, 1990.

[4] Huang, X., A. Acero, H.-W. Hon, and R. Reddy: *Spoken language processing: A guide to theory, algorithm, and system development*, vol. 95. Prentice Hall PTR Upper Saddle River, 2001.

[5] Adami, A. G., R. Mihaescu, D. A. Reynolds, and J. J. Godfrey: *Modeling prosodic dynamics for speaker recognition*. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. IV–788. 2003.

[6] Dehak, N., P. Dumouchel, and P. Kenny: *Modeling prosodic features with joint factor analysis for speaker verification*. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7), pp. 2095–2103, 2007.

[7] Murty, K. S. R. and B. Yegnanarayana: *Combining evidence from residual phase and MFCC features for speaker recognition*. *IEEE Signal Processing Letters*, 13(1), pp. 52–55, 2005.

[8] KINNUNEN, T. and P. ALKU: *On separating glottal source and vocal tract information in telephony speaker verification.* In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4545–4548. 2009.

[9] TITZE, I. R. and B. H. STORY: *Acoustic interactions of the voice source with the lower vocal tract. The Journal of the Acoustical Society of America*, 101(4), pp. 2234–2243, 1997.

[10] HONDA, K., T. KITAMURA, H. TAKEMOTO, S. ADACHI, P. MOKHTARI, S. TAKANO, Y. NOTA, H. HIRATA, I. FUJIMOTO, Y. SHIMADA ET AL.: *Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling. Computer Methods in Biomechanics and Biomedical Engineering*, 13(4), pp. 443–453, 2010.

[11] GODOY, E., N. MALYSKA, and T. F. QUATIERI: *Estimating lower vocal tract features with closed-open phase spectral analyses.* In *Proc. of the Interspeech 2015*. Dresden, Germany, 2015.

[12] GODOY, E., A. DUMAS, J. MELOT, N. MALYSKA, and T. F. QUATIERI: *Relating estimated cyclic spectral peak frequency to measured epilarynx length using magnetic resonance imaging.* In *Proc. of the Interspeech 2016*, pp. 948–952. San Francisco, USA, 2016.

[13] DEHAK, N., 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*, 19(4), pp. 788–798, 2010.

[14] DRUGMAN, T., M. THOMAS, J. GUDNASON, P. NAYLOR, and T. DUTOIT: *Detection of glottal closure instants from speech signals: A quantitative review. IEEE Transactions on Audio, Speech, and Language Processing*, 20(3), pp. 994–1006, 2011.

[15] CHILDERS, D., A. SMITH, and G. MOORE: *Relationships between electroglottograph, speech, and vocal cord contact. Folia Phoniatrica et Logopaedica*, 36(3), pp. 105–118, 1984.

[16] WINKLER, R.: *Merkmale junger und alter Stimmen: Analyse ausgewählter Parameter im Kontext von Wahrnehmung und Klassifikation*, vol. 6. Logos Verlag Berlin GmbH, 2008.

[17] ELLBOGEN, T., F. SCHIEL, and A. STEFFEN: *The BITS speech synthesis corpus for German. Age*, 47(45), p. 40, 2004.

[18] KADIRI, S. R., R. PRASAD, and B. YEGNANARAYANA: *Detection of glottal closure instant and glottal open region from speech signals using spectral flatness measure. Speech Communication*, 116, pp. 30–43, 2020.

[19] PRATHOSH, A., T. ANANTHAPADMANABHA, and A. RAMAKRISHNAN: *Epoch extraction based on integrated linear prediction residual using plosion index. IEEE Transactions on Audio, Speech, and Language Processing*, 21(12), pp. 2471–2480, 2013.

[20] REYNOLDS, D. A., T. F. QUATIERI, and R. B. DUNN: *Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing*, 10(1-3), pp. 19–41, 2000.

[21] JAWARKAR, N. P., R. S. HOLAMBE, and T. K. BASU: *Effect of nonlinear compression function on the performance of the speaker identification system under noisy conditions.* In *Proc. of the 2nd International Conference on Perception and Machine Intelligence*, pp. 137–144. 2015.

[22] DAQROUQ, K. and T. A. TUTUNJI: *Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers*. *Applied Soft Computing*, 27, pp. 231–239, 2015.

[23] NASSIF, A. B., I. SHAHIN, S. HAMSA, N. NEMMOUR, and K. HIROSE: *CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions*. *Applied Soft Computing*, 103, p. 107141, 2021.