

On the Potential of Glottal Signatures for Speaker Recognition

Thomas Drugman, Thierry Dutoit

TCTS Lab, University of Mons, Belgium

Abstract

Most of current speaker recognition systems are based on features extracted from the magnitude spectrum of speech. However the excitation signal produced by the glottis is expected to convey complementary relevant information about the speaker identity. This paper explores the use of two proposed glottal signatures, derived from the residual signal, for speaker identification. Experiments using these signatures are performed on both TIMIT and YOHO databases. Promising results are shown to outperform other approaches based on glottal features. Besides it is highlighted that the signatures can be used for text-independent speaker recognition and that only several seconds of voiced speech are sufficient for estimating them reliably.

Index Terms: Speaker Recognition, Glottal Analysis, Residual Signal, Glottal Signature, Voiceprint

1. Introduction

Building an efficient speaker recognition system implies to have a good understanding of what defines the speaker individuality. Although high-level information (such as the word usage) could be of interest, low-level acoustic features are generally employed [1]. These latter features are most of the time extracted from the amplitude spectrum of the speech signal. They aim at parametrizing the contribution of the vocal tract, which is an important characteristic of the speaker identity. On the other hand, very few works address the possibility of using features derived from the glottal source in speaker recognition. However significant differences in the glottal waveforms have been observed between different speaker types [2].

Two main signals convey information about the glottis behaviour: the glottal flow and the residual signal. The glottal flow is the airflow arising from the trachea and passing through the vocal folds. Its estimation directly from speech signals is a typical problem of blind separation since neither the glottal nor the vocal tract contributions are observable. It is then required to adopt an estimation process incorporating a deep knowledge of the production mechanism. In this way, the glottal flow can be estimated for example by a closed-phase linear predictive analysis. Using this technique, Plumpe et al. [3] extracted a set of time features parametrizing the estimated glottal flow. In a similar framework, Gudnason et al. [4] characterized the glottal flow by real cepstrum coefficients. These two approaches reported an improvement in terms of speaker identification when combining glottal parameters to features extracted from the amplitude spectrum of speech (such as LP or MFCC coefficients). As for the residual signal, it refers to the signal obtained by inverse filtering, after removing the spectral envelope contribution. The resulting residual signal then conveys relevant information about the excitation and, contrarily to the glottal flow, has the advantage of being easily obtained. In [5], Thevenaz et al. suggested to use the LPC coefficients of the residual signal

in speaker verification. More recently, Murty et al. [6] emphasized the complementarity of the residual phase with conventional MFCCs in speaker recognition. In that study, the information contained in the residual phase was extracted by using bottleneck neural networks.

The goal of this paper is to investigate the potential of using *glottal signatures* in speaker recognition. The research of an invariant *voiceprint* in the speech signal, univoquely characterizing a person (as achieved with the fingerprint), has always attracted the speech community [7]. As this seems utopian due to the inherent nature of the phonation mechanism, we here prefer the use of the term "*signature*" for denoting a signal conveying a relevant amount of information about the speaker identity.

The paper is structured as follows. In Section 2, it is detailed how these signatures are extracted from the residual signal. It is also given more detail on the amount of data required for reliably estimating them and it is shown that they can be used for text-independent speaker recognition. Section 3 presents experiments of speaker identification led on both TIMIT and YOHO databases. Finally Section 4 concludes.

2. Glottal Signatures

2.1. Glottal Signatures used in this Study

The so-called *signatures* used in this study are derived from the Deterministic plus Stochastic Model (DSM) of the residual signal we proposed in [8] for parametric speech synthesis. This model arises from an analysis led on a dataset of pitch-synchronous normalized residual frames. The workflow for obtaining this particular dataset from a collection of speaker-dependent recordings is presented in Figure 1. First a traditional Linear Prediction (LP) analysis capturing the spectral envelope is performed on the speech signals. Residuals are then obtained by inverse filtering. Glottal Closure Instants (GCIs) are then identified by locating the greatest discontinuities in the residual signal as explained in [9]. In parallel, the pitch is estimated using the publicly available Snack Sound Toolkit [10]. The pitch-synchronous residual frames are then isolated by a GCI-centered, two pitch period-long Blackman windowing. The resulting frames are finally normalized in prosody, i.e normalized both in pitch and energy. The pitch-normalization operation is achieved by decimation/interpolation on a given fixed number of samples (so that all residual frames have the same length).

Once the dataset of residual frames is available, some speaker-dependent features related to the DSM are extracted from it. According to this model [8], the voiced residual signal $r(t)$ is composed of a low-frequency deterministic structure $r_d(t)$ and a high-frequency stochastic component $r_s(t)$, assumed to principally model the turbulences of the glottal airflow. The spectrum is then divided into two bands delimited by the so-called *maximum voiced frequency* F_m (fixed to $4k\text{Hz}$ in the present study). The synthesized residual signal is then obtained as described in Figure 2. The deterministic part is

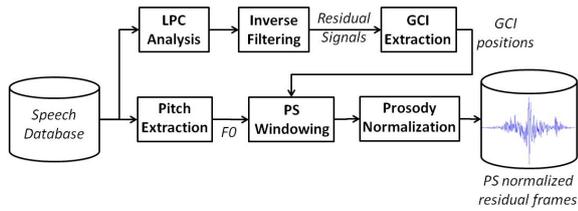


Figure 1: Workflow for obtaining the speaker-dependent dataset of pitch-synchronous normalized residual frames.

modeled by a single speaker-dependent waveform called *first eigenresidual*. This waveform is defined as the first eigenvector obtained by computing a Principal Component Analysis (PCA) on the dataset of residual frames. As for the stochastic component, it is modeled by a high-frequency Gaussian noise modulated in time by a pitch-synchronous energy envelope. This *energy envelope* is derived from the previous dataset by averaging the Hilbert envelope of the high-frequency contents of the residual frames.

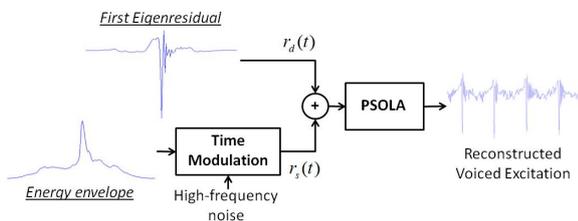


Figure 2: Voiced excitation reconstruction according to the Deterministic plus Stochastic Model (DSM) of the residual signal. The two signatures used in this work are the first eigenresidual and the energy envelope.

As a conclusion, the DSM of the residual signal makes use of two speaker-dependent waveforms, hereafter called *glottal signatures*: the first eigenresidual (or eigenresidual for short) and the noise energy envelope. Figure 3 displays the shape of the energy envelope for two male speakers. Differences in the waveforms suggest that the proposed glottal signatures have the potential to be used for automatic speaker recognition.

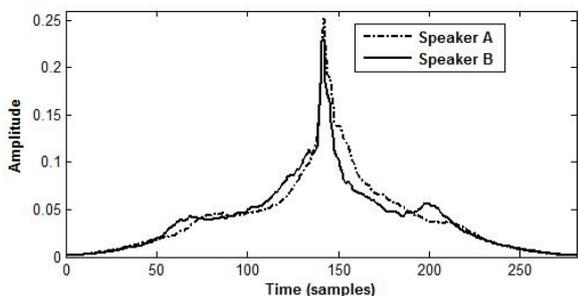


Figure 3: Waveform of the energy envelope for two different male speakers.

2.2. Using the Glottal Signatures in Speaker Identification

So as to be integrated into a speaker identification system, the signatures are estimated on both training and testing sets. A

confusion matrix $C(i, j)$ between speaker i and speaker j is then computed. In this work, the Relative Time Squared Error (RTSE) was chosen as a distance measure between two different waveforms. If $v_{k,l,training}$ and $v_{k,l,test}$ denote the k^{th} signature (in our case, $k = 1, 2$ respectively for the eigenresidual and the energy envelope) for speaker l estimated respectively on the training and testing datasets, the confusion matrix $C_k(i, j)$ using only the k^{th} signature is defined as:

$$C_k(i, j) = \sqrt{\frac{\sum_{n=0}^{N-1} (v_{k,i,test}(n) - v_{k,j,training}(n))^2}{\sum_{n=0}^{N-1} v_{k,j,training}(n)^2}} \quad (1)$$

where N is the number of samples used for the pitch normalization. The confusion matrix $C(i, j)$ is finally obtained as:

$$C(i, j) = C_1(i, j) \cdot C_2(i, j) \quad (2)$$

Note that several operations for combining the two matrices are possible. Among our experiments, the multiplication gave the best results, although the differences we observed were relatively weak.

Finally, the identification of a speaker i is carried out by looking for the lowest value in the i^{th} row of the confusion matrix $C(i, j)$. The speaker is then correctly identified if the position of the minimum is i . In other words, when recordings are presented to the system, the identified speaker is the one whose signatures are the closest (in the Euclidian sense) to the signatures extracted on these recordings.

2.3. Speed of Convergence

It can be wondered how much data is required for having a reliable estimation of the two signatures presented in the previous Section. To answer this question, the speaker AWB from the CMU ARCTIC database [11] is analyzed. This database contains about 50 minutes of speech recorded for Text-to-Speech purpose. Reference signatures were first computed on a large dataset containing about 150.000 pitch-synchronous residual frames. In a second time an estimation of these signatures was obtained by repeating the same operation on a held-out dataset for the same speaker. The Relative Time Squared Error (RTSE) is used for both signals as a distance between the estimation and the reference. Figure 4 displays the evolution of this measure with the size of the held-out dataset. It may be observed that both estimations quickly converge towards the reference signatures. From this graph, it can be considered that a dataset containing around 1000 residual frames is sufficient for leading to a correct estimation of both signatures. This corresponds to about 7s of voiced speech for a male speaker, and about 4 s for a female voice.

2.4. Phonetic Class Independence

One may also wonder whether the assumption of using a single waveform for the excitation of all voiced phonetic classes is respected. For this, the same database was first segmented, as classically achieved in statistical parametric speech synthesis [12]. More precisely, spectral parameters were extracted from the database and a 5-state left-to-right HMM-based filter model based on this representation was trained. For each state, a decision tree was built according to phonetic criteria so that each leaf contains at least 1000 frames. This ensures that class-dependent signatures can be reliably estimated within each voiced leaf (see Section 2.3). Through our observations we

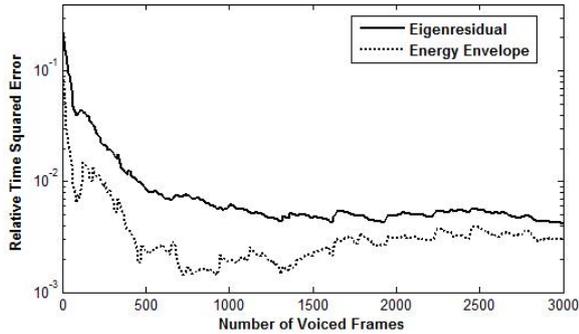


Figure 4: Speed of convergence for the 2 signatures.

reported that these waveforms were fundamentally comparable to each other (with a RTSE generally below 1%). In the context of the present study, this implies that the proposed signatures can be used for *text-independent* (and possibly language-independent) speaker recognition.

3. Experiments

Experiments detailed in this Section were led on both TIMIT and YOHO databases, for comparison purpose with [3] and [4]. The TIMIT database [13] comprises 10 recordings from 630 speakers (438 males, 192 females) and sampled at 16 kHz. As for the YOHO database [14], it contains speech from 138 speakers (108 males, 30 females) sampled at 8 kHz. These recordings were collected in a real-world office environment through 4 sessions over a 3 month period. For each session, 24 phrases were uttered by each speaker. In these experiments, the data was split for each speaker (and each session for YOHO) into 2 equal parts for the training and the testing. This is done in order to guarantee that, for both steps, enough residual frames are available for reliably estimating the signatures (see Section 2.3).

3.1. Results on the TIMIT database

To give a first idea of the potential of using the glottal signatures in speaker recognition, Figure 5 displays the distributions of $C_1(i, j)$ respectively when $i = j$ and when $i \neq j$. In other words, this plot shows the histograms of the RTSE (in logarithmic scale) between the eigenresiduals estimated respectively for the same speaker and for different speakers. It is clearly observed that the error measure is much higher (about 15x in average) when the tested signature does not belong to the considered speaker. It is also noticed that, for the same speaker, the RTSE on the eigenresidual is about 1%, which confirms our results of Sections 2.3 and 2.4. However a weak overlap between both distributions is noted, which may lead to some errors of speaker identification.

Figure 6 exhibits the evolution of the identification rate with the number of speakers considered in the database. For this, the identification was achieved using only one of the two glottal signatures, or using their combination as suggested in Equation 2. As expected the performance degrades as the number of speakers increases, since the risk of confusion becomes more important. However this degradation is relatively slow in all cases. One other important observation is the clear advantage of combining the information of the two signatures. Indeed this leads to an improvement of 7.78% compared to using only the eigenresidual.

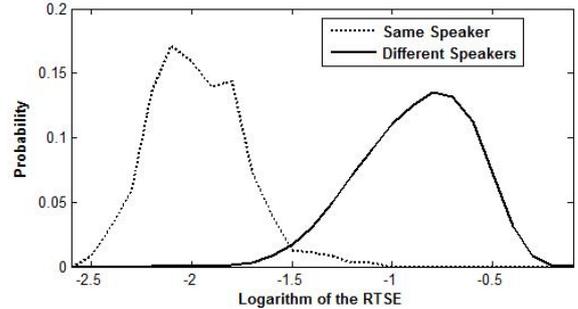


Figure 5: Distributions of the Relative Time Squared Error (RTSE) between the eigenresiduals estimated respectively for the same speaker and for different speakers.

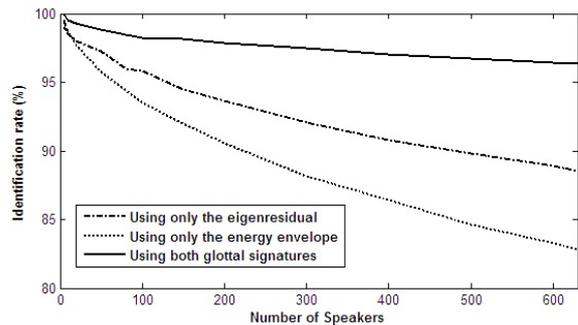


Figure 6: Evolution of the identification rate with the number of speakers for the TIMIT database.

Table 1 summarizes the results obtained on the TIMIT database. Identification rates for 168 speakers are also given for comparison purpose. Indeed in [3] Plumpe et al. extracted a set of 12 time-domain parameters characterizing the glottal flow estimated by a closed-phase analysis. Using these features they reported an average misclassification rate of 28.64% using a subset of 168 speakers. On the same subset, Gudnason et al. reported in [4] a misclassification rate of 5.06% using voice source cepstrum coefficients. These results can be compared to the 1.98% we achieved using the two signatures. Finally note that Gudnason et al. [4], using the glottal features, also obtained a misidentification rate of 12.95% on the whole TIMIT database (630 speakers). With the proposed signatures, a misclassification rate of 3.65% is reached. It is worth noting that we observed that the proposed approach works equivalently for both male and female speakers.

	168 speakers	630 speakers
Using only the eigenresidual	5.88	11.43
Using only the energy envelope	8.76	17.14
Using both glottal signatures	1.98	3.65

Table 1: Misidentification rate (%) on the TIMIT database obtained using only one glottal signature or both of them.

3.2. Results on the YOHO database

Compared to the TIMIT database, the YOHO corpus differs in two main aspects: 1) recordings are now sampled at 8 kHz, 2) recordings were collected in several sessions over a period

of 3 months. The first point implies that only the eigenresidual will be used here (since F_m is fixed to 4 kHz, and hence only the deterministic part is here considered). Regarding the second aspect, one can expect a higher intra-speaker variability when training and testing sessions are spaced over a long period of time. Results we obtained on the YOHO corpus using both signatures are presented in Figure 7. These results are detailed according the period separating the training and test recordings. Besides the percentages of cases for which the correct speaker is recognized in second or third position (instead of first position) are also given. From this plot it can be noted that the system works perfectly when recordings are from the same session. Compared to results from the TIMIT database (where about 95% were obtained for 138 speakers with the eigenresidual), the performance is higher here. This is among others due to the greatest amount of data for extracting the signature. On the contrary, when the test is done one session later, the identification dramatically drops till 70%. This first fall is mainly due to the mismatch between training and testing conditions. It is observed that the identification rate then decreases of about 5% for any later session. As expected, this results from the higher speaker variability as the time interval between sessions increases. Note also that, when training and testing conditions differ, between 12% and 16% of speakers are identified in second or third position. It can then be expected that combining the proposed glottal signatures with features based on the speech magnitude spectrum will remove most of this ambiguity. Finally, for comparison purpose, Gudnason et al. reported in [4] a misclassification rate of 36.3% using voice source cepstral coefficients (with test recordings coming from the 4 sessions). Averaging our results over all the sessions, we found a misclassification rate of 29.3% using only the eigenresidual.

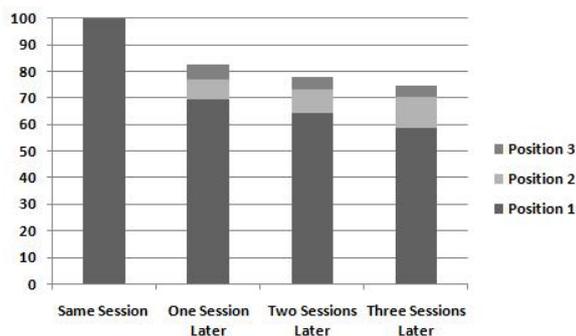


Figure 7: Identification rate (%) for the YOHO database when training and testing sessions may be separated over a long period. The proportion of speakers for which the correct signature is recognized in second or third position is also given.

4. Conclusion

This paper investigated the potential of using glottal signatures for speaker recognition. These signatures were derived from an analysis of a speaker-dependent dataset of pitch-synchronous prosody-normalized residual frames. It was shown that about 1000 voiced frames are sufficient for a reliable estimation of both signatures. Besides the estimation was also shown to converge towards the same waveform independently of the considered phonetic class. The signatures could then be used in text-independent speaker recognition. Results of speaker identification were reported on both TIMIT and YOHO databases. In

these experiments, the proposed signatures outperformed other similar studies based on glottal features. However the performance was shown to degrade when training and testing sessions are spaced in time.

Several improvements could be brought to the current approach. Indeed results were obtained using *only* the proposed glottal signatures. Regarding the evidence of a complementarity between excitation-based and vocal tract-based features ([6], [3], [4]), it is reasonable to expect that combining the proposed signatures with a conventional speaker recognition system (e.g. with a typical GMM-MFCC approach) would lead to an appreciable improvement. Secondly, applying some channel compensation could alleviate the mismatch between training and testing sessions. Indeed different recording conditions impose different characteristics to the speech signal. Among these, differences in phase response may dramatically affect the estimation of the signatures (since the information of the residual is essentially contained in its phase). These two possible improvements are the object of ongoing work.

5. Acknowledgments

Thomas Drugman is supported by the “Fonds National de la Recherche Scientifique” (FNRS).

6. References

- [1] D.A. Reynolds, *An overview of automatic speaker recognition technology*, Proc. ICASSP, vol. 4, pp. 4072-4075, 2002.
- [2] I. Karlsson, *Glottal Waveform Parameters for Different Speaker Types*, STL-QPSR, vol. 29, pp. 6167, 1988.
- [3] M. Plumpe, T. Quatieri, D. Reynolds, *Modeling of the glottal flow derivative waveform with application to speaker identification*, IEEE Trans. on Speech and Audio Processing, vol. 7, pp. 569-586, 1999.
- [4] J. Gudnason, M. Brookes, *Voice source cepstrum coefficients for speaker identification*, Proc. ICASSP, pp. 4821-4824, 2008.
- [5] P. Thevenaz, H. Hugli, *Usefulness of the LPC-residue in text-independent speaker verification*, Speech Communication, vol. 17, pp 145-157, 1995.
- [6] S. Murty, B. Yegnanarayana, *Combining evidence from residual phase and MFCC features for speaker recognition*, IEEE Signal Processing Letters, vol. 13, pp. 52-55, 2006.
- [7] L.G. Kersta, *Voiceprint Identification*, Nature 196, pp. 1253-57, 1962.
- [8] T. Drugman, G. Wilfart, T. Dutoit, *A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis*, Proc. Interspeech, 2009.
- [9] T. Drugman, T. Dutoit, *Glottal Closure and Opening Instant Detection from Speech Signals*, Interspeech Conference, 2009.
- [10] [Online], “The Snack Sound Toolkit”, <http://www.speech.kth.se/snack/>.
- [11] [Online], “CMU ARCTIC speech synthesis databases”, <http://festvox.org/cmu-arctic/>.
- [12] H. Zen, K. Tokuda, A. Black, *Statistical Parametric Speech Synthesis*, Speech Communication, vol. 51, pp 1039-1064, 2009.
- [13] W. Fisher, G. Doddington, K. Goudie-Marshall, *The DARPA Speech Recognition Research Database: Specifications and Status*, Proc. DARPA Workshop on Speech Recognition, pp. 9399, 1986.
- [14] J. Campbell, *Testing with the YOHO CD-ROM Voice Verification Corpus*, Proc. ICASSP, pp. 341344, 1995.