2010 International Symposium on Nonlinear Theory and its Applications
NOLTA2010, Krakow, Poland, September 5-8, 2010

Speaker Identification with Voiced Speech Variability Modeling using Phase Space Reconstruction

Łukasz Bronakowski and Krzysztof Ślot

Technical University of Lodz, Institute of Electronics
ul. Wolczanska 211/215, Lodz, Poland
Email: lukasz.bronakowski@p.lodz.pl, kslot@p.lodz.pl

Abstract—The following paper examines a possibility of applying concepts and methods of chaotic system analysis for speech variability modeling in speaker identification task. The proposed descriptor comprise a set of parameters, which are derived for reconstructed phase spaces of voiced-speech segments. The proposed method for analysis of attractor convergence is based on correlation sum vectors, summarised by vector quantization technique. It has been shown that the presented approach appears to be promising means for speaker discrimination.

1. Introduction

In recent years one can observe an increasing interest in exploiting new features that can improve performance of biometric identity verification systems. Speaker recognition is one of the most prominent research directions within this field, as it pertains to speech, which is the basic behavioral biometric modality. Majority of commonly used methods in speaker recognition that have been proposed so far exploit short- and long-term spectral and energy features, such as Linear Prediction Coefficients - LPC, Mel-Frequency Cepstral Coefficients - MFCC and corresponding regression coefficients (delta-cepstral, delta-delta-cepstral coefficients) [1]. These features reflect only the behavioral aspect of speech production, which comprises processes involved in phone articulation.

Majority of vocal tract components are highly controllable, yielding a wide variability of possible ways of speech articulation, which is a drawback from the point of view of biometrics. However, a voluntary control over vocal tract is only partial in case of vocal folds and their excitation mechanisms, which are strongly determined by anatomical factors. Therefore, these components of the speech production system seem to naturally fit biometric purposes. Among the pool of features that are widely used in research on speaker recognition a characteristic that is directly related to vocal fold operations is the fundamental frequency of speech. However, this descriptor provides information on the speech-production aspect, which can be voluntarily controlled, revealing no evidence on more complex speech-phonation mechanisms. The phonation process, which is determined by anatomic structure of vocal folds and glottis, is generally omitted in speaker recognition methods. The reason for it could be the difficulty and complexity of speech production, which involves aerodynamic, biomechanical, and acoustic factors that are still not fully understood. One of the open issues in speech signal exploration is a phenomenon of short-time variability in voiced speech production.

This phenomena can be interpreted as bifurcations and low-dimensional chaos [2], therefore non-linear theory can be applied to perform discussed problem. Descriptors of nonlinear speech behavior that account for vocal folds individual anatomy could provide a promising basis for identity resolution, as physiological features are known to be the most reliable biometric characteristics (as it is in case of retina, iris or fingerprints).

An objective of the following paper is to examine, whether short-time variability in speech-production can be exploited as a useful feature for speaker discrimination. We hypothesize that the speech-phonation aspect, reflected by speech signal non-stationary behavior, can be as important in speaker-modeling as the commonly used spectral characteristics of the human vocal tract. To verify the formulated hypothesis we propose a novel speech signal descriptor: a measure of convergence of reconstructed phase space attractors, derived for voiced speech segments.

The organization of the paper is as follows: the proposed descriptor of speech signal variability is introduced in section 2; speaker identification procedure is outlined in section 3 and its experimental evaluation is discussed in section 4.

2. Voiced-speech Variability Descriptor

Due to a limited rate of speech organ dynamics, one can assume that the speech signal is stationary within approximately 30 ms intervals. Small signal perturbations, at the order of no more than one percent, which are always present over this quasi-stationary background, heavily contribute to an individual appearance of speech. Furthermore, these perturbations can provide information about psycho-physiological state of the speaker, because they can originate from changes in tension of articulatory apparatus muscles and fluctiations of the air pressure exhaled from lungs during speaking [2].

Short-term speech variability is involuntary - one cannot

control a vocal folds tension at millisecond-long rates (as opposed to intentional control over changes of fundamental frequency over long-term intervals). As a result, it is an invaluable source of information on physical rather than behavioral properties of vocal tract, which can be exploited for identity recognition.

Speech signal analysis tools that have been used throughout the reported research for examining short-term perturbations have been adopted from nonlinear system theory and include Poincare mapping, reconstructed phase space and fractal dimension analysis. High speaker-discrimination potential that is offered by the presented signal description perspective can be easily noticed from plots presented in Fig. 1, where reconstructed phase space is used to present the same utterance (a vowel 'a') spoken by three different speakers.

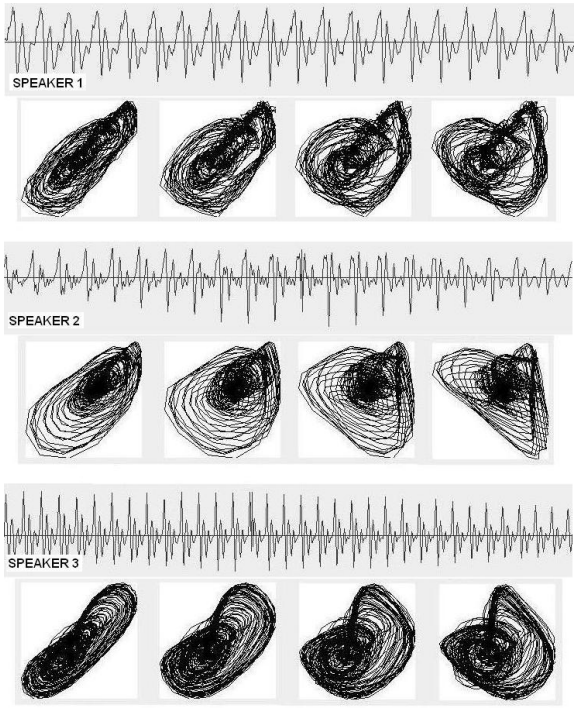

Figure 1: Speech waveform of the vovel /a/ spoken by three different speakers and corresponding reconstructed phase spaces obtained from four consecutive speech frames.

2.1. Reconstructed Phase Space

The reconstructed phase space (RPS) can be considered as a plot of the time-lagged version of a signal. Structural patterns that occur in such phase space are commonly reffered to as trajectories or attractors.

For time series $x(n)$, where $n = 1...N$ is a time index, each RPS trajectory point is a vector [3]:

$$\mathbf{x}_n = \begin{bmatrix} x_n & x_{n-\tau} & \cdots & x_{n-(d-1)\tau} \end{bmatrix} \quad (1)$$

where τ is a time lag and d is an embedding dimension. The RPS trajectory of the whole signal can be presented as a matrix composed of time-delayed vectors $\mathbf{x}_n$:

$$\mathbf{X} = \begin{bmatrix} x_{1+(d-1)\tau} & \cdots & x_{1+\tau} & x_1 \\ x_{2+(d-1)\tau} & \cdots & x_{2+\tau} & x_2 \\ \vdots & & \ddots & \\ x_N & \cdots & x_{N-(d-2)\tau} & x_{N-(d-1)\tau} \end{bmatrix} \quad (2)$$

RPS representation of a signal captures full dynamics of the underlying system and includes nonlinear information, which is not preserved by commonly used in speaker-recognition spectral-based speech-representation techniques. Several methods have been developed for estimation of phase-space trajectory distribution. These are e.g. Bayesian modeling of scatter of samples [4] or correlation dimension [5]. In the reported research, the correlation sum have been adopted as a statistical descriptor of the underlying attractor [2]. Correlation sum measures a trajectory divergence rate and is given by:

$$C(\varepsilon) \sim \sum_{ij} \Theta(\varepsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \quad (3)$$

where $\Theta(x)$ is a Heaviside function and sum indices refer to trajectory points subject to testing (i) and points of their neighborhoods (j). To get a more complete information on a structure of an attractor, we propose to characterize each frame m of an input signal using the following descriptor:

$$\mathbf{C}^m = [C(\varepsilon_1)^m, C(\varepsilon_2)^m ... C(\varepsilon_p)^m] \quad (4)$$

where p is the adopted maximum size of neighborhood of interest and $C(\varepsilon_k)^m$ is given by:

$$C(\varepsilon_k)^m = \frac{1}{MN} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ i \neq j}}^{N} \Theta(\varepsilon_k - \|\mathbf{x}_i - \mathbf{x}_j\|) \quad (5)$$

The sum $C(\varepsilon_k)^m$ is computed at M trajectory samples per frame and N is the total number of trajectory points. The procedure of attractor descriptor derivation has been schematically depicted in Fig. 2. As a result of its application is derivation of a $p-$element vector (4), which will be used for speech signal frame representation in subsequent classification.

3. Speaker Identification Procedure

Data classification methodology that has been adopted for recognition (presented schematically in Fig. 3) assumes no temporal ordering of frames, which is a common approach for text-independent speaker recognition tasks . All vectors (4) extracted from frames of input training sequences (voiced-speech segments of sentences) uttered by a given speaker are used to build a reference model for this speaker. The model is a set of q-vectors that are codebook elements derived from vector-quantization of the corresponding distribution of training samples.

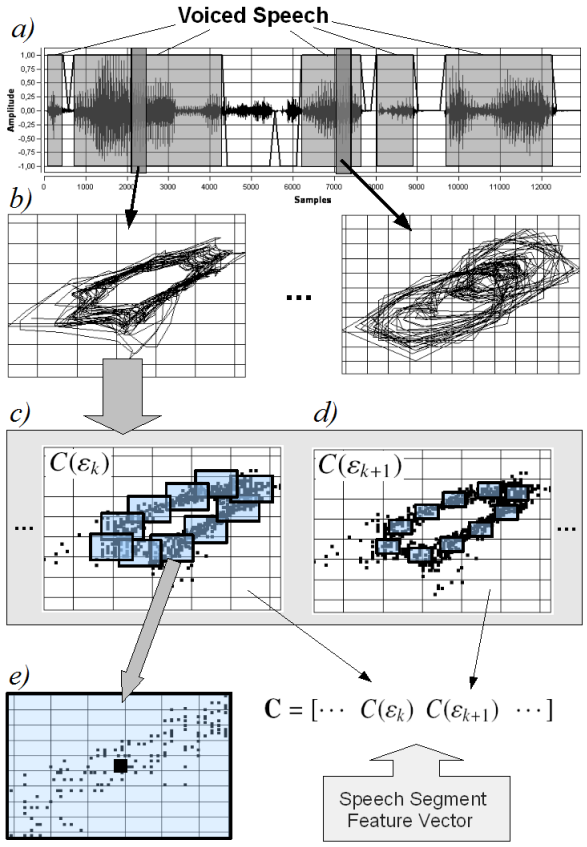

Figure 2: Schematic diagram of a proposed descriptor derivation procedure: input speech (a), RPS representation of two sample frames (b), an RPS plot for two levels of detail in calculating of correlation sums (c and d) and sample distribution of points in a neighborhood ε_k of a trajectory point (e).

Each speaker s is therefore represented by a unique set of q-vectors $\{\mathbf{v}_i^s\}$. Speaker recognition is a procedure of confronting subsets of L test vectors $\{\mathbf{c}_i\}$ extracted from consecutive frames of a voiced-speech segment, with all available models (codebooks). A measure of fit is an overall distortion between codebook vectors of a given model and the test sequence vectors:

$$D_s = \frac{1}{L} \sum_{i=1}^{L} d(\tilde{\mathbf{v}}_i^s, \mathbf{c}_i) \tag{6}$$

The vector $\tilde{\mathbf{v}}_i^s$ is the closest match between the test vector c_i and all codebook elements from a model of the considered speaker s:

$$\tilde{\mathbf{v}}_i^s = \min_{1 \le j \le q} d(\mathbf{v}_j^s, \mathbf{c}_i) \tag{7}$$

Euclidean distance has been used as a distance measure $d(.,.)$ throughout the reported research.

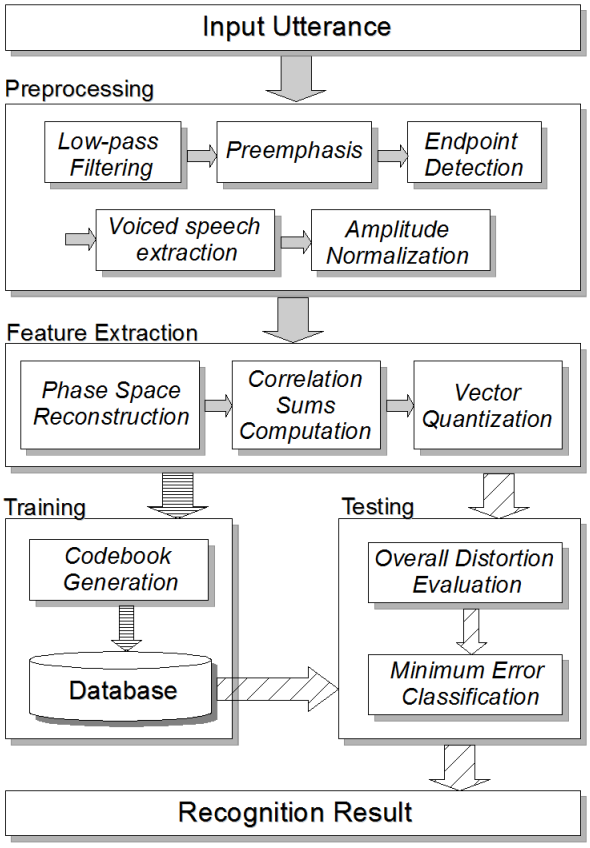

Figure 3: Speaker identification procedure.

4. Experimental Evaluation of the Proposed Procedure

The presented procedure has been verified using publicly available database CSLU: Speaker Recognition Version 1.1 [6]. Five different sentences, repeated twice by ten speakers (5 females and 5 males) during twelve recording sessions were subject to analysis. Each of the sentences comprised several voiced-speech segments, yielding sets of about 14000 voiced phonemes (frame sequences) per speaker. This set was evenly split into training and testing parts.

Pronunciation variability has been assessed for speech segments of about 30 ms length with 10 ms overlap. The time lag τ used for generate RPS is in general empirical, but we adopted a commonly used measure based on autocorrelation function: $C(\tau) = 0.5 \cdot C(0)$. The embedding dimension d is estimated using the false neighbours method [7]. Eight values of the neighborhood size ε were used for frame-descriptor construction: $0.001, 0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128$ (the magnitude of input signal was normalized within the range [-1 ... 1]).

A value of $M = 10$ (equation (5)) has been arbitrarily adopted, which means that 10 points are selected along a trajectory as reference points for correlation sum computation. The points are selected in such a way that each of the

resulting between-point intervals contains approximately 10% of data points. To determine length of the trajectories, an estimation of the fundamental frequency is made according to the algorithm which is based on the computation of autocorrelation of speech in time-domain [8].

K-means method [9] is used to generate codebooks (speaker models) from correlation sum vectors. The number of means have been varied between 8 and 128. Speaker recognition performance shown as a function of a codebook-size (separately for male and female speakers) has been presented in Fig. 4. As it can be seen, the best results - over 80% correct recognition - have been obtained for a 32-element codebook. We consider this result to be a very good one, as we apply for speaker recognition a completely different basis than commonly used cepstral coefficients. For example, a use of MFCCs features in combination with GMM modeling and SVM classification [10] yields recognition rates between 66,37% (for 10s test and training speech duration) and 91,87% (for 10s test and 24min training speech duration) for the same CSLU database as in presented paper.

Speech signal descriptors that have been used in the reported research have little in common with spectral characteristics that dominate current techniques. As such, there exist an expectation to sustantially increase speaker-recognition performance if both diverse signal analysis directions are appropriately combined.

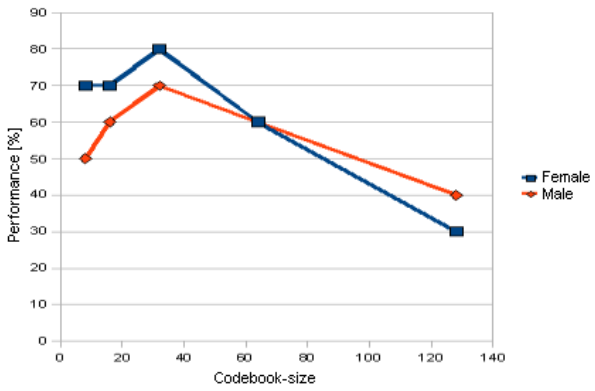

Figure 4: Speaker-recognition performance, shown as a function of codebook-size.

5. Conclusion

The presented paper shows that short-time variability of speech is a source of important clues for speaker recognition. Although the analysis of the problem is still in its early stage - the variability descriptor can be certainly refined - quite good recognition rates can be attained.

The adopted approach to speaker recognition is complementary to the commonly used strategies. Therefore, a natural way of continuation of the reported research is to combine short-term variability with conventional ways of speech-signal characterization and to verify, whether such a combination could result in noticeable improvement in speaker recognition performance.

Acknowledgments

This research has been supported by Polish Ministry of Science and Higher Education under a grant no. N N515 343336.

References

[1] S. Furui, “Recent advances in speaker recognition”, Pattern Recognition Letters 18, 859-872, Elsevier Ltd., 1997

[2] H. Herzel, “Non-Linear Dynamics of Voiced Speech”, in Nonlinear Dynamics: New Theoretical and Applied Results, J. Awrejcewicz eds., Akad. Verl., 1995

[3] R.J. Povinelli, M.T. Johnson, A.C. Lindgren, F.M. Roberts and Y. Jinjin, “Statistical models of reconstructed phase spaces for signal classification”, IEEE Trans. on Sig. Proc., Vol. 54, No. 6, pp.2178-2186, 2006

[4] J. Ye, M. T. Johnson and R. J. Povinelli, “Study of attractor variation in the reconstructed phase space of speech signals”, in Proc. ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP), Le Croisic, pp.5-10, 2003

[5] J. J. Jiang, Yu Zhang, “Nonlinear dynamic analysis of speech from pathological subiects”, Electronics Letters, Vol. 38, No. 6, pp. 294-295, 2002

[6] CSLU: Speaker Recognition Version 1.1, Linguistic Data Consortium, Philadelphia, 2006, LDC Catalog No.: LDC2006S26

[7] L. Zhao, L. Wang, Y. Wang, Z. Huang and N. Fang, “Determining Minimum Embedding Dimension in Short Time Series using Precisely Averaged False-nearest-neighbors Approach”, Microwave Conference, China-Japan, pp.554-557, 2008

[8] L. Hui, B. Dai and L. Wei, “A Pitch Detection Algorithm Based on AMDF and ACF”, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. I-277 – I-380, 2006

[9] J. R. Deller Jr., J. H. L. Hansen, J. G. Proakis, “Discrete-Time Processing of Speech Signals”, Wiley-IEEE Press, 1999

[10] E. Dikici, M. Saralar, “Investigating the Effect of Training Data Partitioning for GMM Supervector Based Speaker Verification”, IEEE 24th International Symposium on Computer and Information Sciences (ISCIS 2009), Northern Cyprus, 2009