DE GRUYTER OPEN

Journal of ELECTRICAL ENGINEERING, VOL. 66, NO. 3, 2015, 169–173

COMMUNICATIONS

TEXT–INDEPENDENT SPEAKER RECOGNITION USING TWO–DIMENSIONAL INFORMATION ENTROPY

Boško Božilović [*] — **Branislav M. Todorović** [**]
Miroslav Obradović [*]

Speaker recognition is the process of automatically recognizing who is speaking on the basis of speaker specific characteristics included in the speech signal. These speaker specific characteristics are called features. Over the past decades, extensive research has been carried out on various possible speech signal features obtained from signal in time or frequency domain. The objective of this paper is to introduce two-dimensional information entropy as a new text-independent speaker recognition feature. Computations are performed in time domain with real numbers exclusively. Experimental results show that the two-dimensional information entropy is a speaker specific characteristic, useful for speaker recognition.

K e y w o r d s: biometrics, speech, speaker recognition, feature extraction, information entropy

1 INTRODUCTION

Biometric recognition systems are increasingly being deployed as a means for the recognition of people [1]. One of the most widely used biometric modalities is human voice. Speaker recognition systems are technologies which are used to recognize person from his/her speech signal by exploiting speaker specific characteristics [2].

Speaker specific characteristics are result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals. In speaker recognition systems, all these speaker specific characteristics can be used to discriminate between speakers [3]. These speaker specific characteristics are called features. The most important characteristic of feature would be large between-speaker variability and small within-speaker variability [4].

Speech signal is a complex time-varying signal which can be represented by many different features. There are different ways to categorize the features. From the viewpoint of their physical interpretation, we can divide them into: spectral features [5, 6], phonetic features [7, 8] and prosodic features [9].

Spectral features are computed from short frames of about 20–30 ms in duration. Within this interval, the speech signal is assumed to remain stationary. Spectral features represent the most common way to characterize the speech signal. Fourier analysis provides a usual way of analyzing the spectral properties of a given signal in the frequency domain. In speech analysis, the phase spectrum is usually neglected, since it is generally believed that it has little effect on the perception of speech [10]. The simplest way of analyzing spectral properties of a signal is by using filter banks. This approach to spectral feature extraction is so called subband filtering where subband outputs are considered directly as the features [11]. The most frequently used spectral features for speaker recognition are mel-frequency cepstral coefficients [12], which are based on mel-scale filter banks. Linear prediction [13, 14] is an alternative spectrum estimation method.

Phonetic features depend on speech content [15]. In order to extract phonetic features it is necessary to perform segmentation of the speech signal into phonemes. Some broad phonetic classes are more speaker specific than others. For example, using only vowels it is possible to obtain a very high recognition rate [16].

Prosodic features are related to non-segmental aspects of speech. They reflect differences in speaking style, language background, sentence type and emotions [17]. The most important prosodic parameter is the fundamental frequency [18]. Other prosodic features for speaker recognition include speaking rate, pause statistics and intonation patterns [19].

Depending on the algorithm used, the process of speaker recognition can be categorized as text-dependent and text-independent. Text-independent recognition is the much more challenging of the two tasks, since in text-independent systems there are no constraints on the words which the speakers are allowed to use.

[*] VLATACOM, R&D Center, Milutina Milankovića 5, 11070 Belgrade, Serbia {Bosko; Miroslav.Obradovic}@vlatacom.com; [**] RT-RK, Institute for Computer Based Systems, Narodnog Fronta 23A, 21000 Novi Sad, Serbia, Branislav.Todorovic@rt-rk.com

DOI: 10.2478/jee-2015-0027, Print ISSN 1335-3632, On-line ISSN 1339-309X

In general, phonetic variability represents an adverse factor to accuracy in text-independent speaker recognition. Another adverse factor in text-independent speaker recognition is modeling the different levels of prosodic information (instantaneous, long-term) to capture speaker specific differences [19]. Beside those, adverse factors in speaker recognition include: differences in recording and transmission conditions, influence of noise environment [20], effect of the orthodontic appliances on spectral properties [21], *etc.*

Speaker recognition process is realized in several steps. The first step is speech signal capture by microphone. The second step assumes extraction of speech segments by removing the silence from the captured speech signal. This step is performed by voice activity detector. The next step is the choice of features that will represent the speech signal. The step which follows is the feature extraction process aiming to compute discriminative speech features suitable for speaker recognition. Furthermore, speaker recognition follows a standard procedure which includes two different tasks: speaker identification and speaker verification. In the speaker identification task, an unknown speaker feature is compared against a database of known speakers, and the best matching speaker is identified. An identity claim is given to the speaker verification task, and the speaker's voice sample is compared against the claimed speaker's voice template. If the similarity degree between the voice sample and the template exceeds a predefined decision threshold, the speaker is recognized, and otherwise rejected [19].

State-of-the-art speaker recognition systems use a number of features in parallel, attempting to cover these different aspects and employing them in a complementary way to achieve more accurate recognition [22].

Information entropy can be useful feature for speaker recognition. In information theory, entropy is defined as a measure of the randomness (uncertainty, information content) of a process. The calculation of the entropy of speech is complex as speech signals simultaneously carry various forms of information: phonemes, topic, intonation signals, accent, speaker voice and speaker stylistics. One can consider the entropy of speech signal at several levels: the entropy of words contained in a sequence of speech, the entropy of intonation and the entropy of speech signal features [23].

Information entropy has already been used for speaker recognition. Empirical entropy was proposed in [24], while approximated cross entropy was analyzed in [25].

In this paper, we propose and analyze so-called two-dimensional information entropy as a new feature domain for text-independent speaker recognition. Algorithm for extraction of two-dimensional information entropy from speech signal is described. Experimental results show that the proposed feature domain can be useful to discriminate between speakers.

2 DESCRIPTION OF TWO–DIMENSIONAL INFORMATION ENTROPY

Speech is made up of about 40 basic acoustic symbols, known as phonemes, which are used to construct words, sentences etc. Speech is an information-rich signal that can be represented in frequency or time domain. All this information is conveyed primarily within the traditional telephone bandwidth of 4 kHz [23].

As a speaker specific characteristic of speech signal we use its amplitude-time trajectory. In order to quantify the information content of speech signal in time domain, we define two-dimensional information entropy of amplitude-time trajectory.

Let us consider the analog speech signal $s(t)$ presented in Fig. 1. Maximum value of the signal is denoted with $S_{\max}$, while the minimum value is denoted with $S_{\min}$. One can notice local maximums and local minimums of the signal amplitude, *ie* the time points $(\dots, t_{i-1}, t_i, t_{i+1}, \dots)$ where the first derivative of the signal is equal to zero.

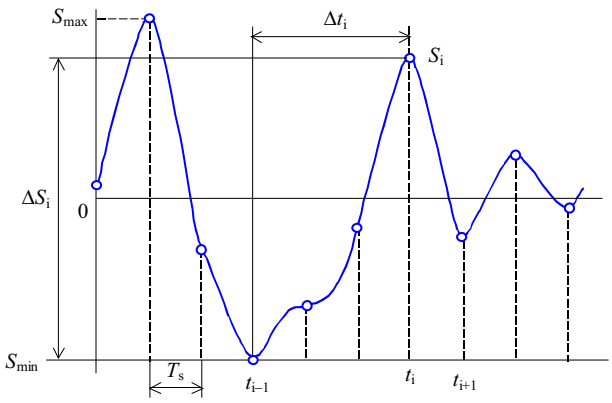

Fig. 1. Speech signal

Let us denote with t_{i-1} time point where the signal has local minimum, with t_i subsequent time point where the signal has local maximum, and with t_{i+1} subsequent time point where the signal has local minimum. Furthermore, let us denote with Δs_i amplitude difference between local maximum at the time point t_i and previous local minimum at the time point t_{i-1}. Let us denote with Δt_i time difference between the time point t_i and the time point t_{i-1}. Similarly, we can define Δs_{i+1} and Δt_{i+1} as amplitude and time differences between time point t_{i+1} and previous time point t_i.

Speech signal is sampled with sampling interval T_s and quantized into q levels. Quantization step is $\Delta q = (S_{\max} - S_{\min})/q$. It should be noted that $\Delta s_i = m\Delta q$ and $\Delta t_i = nT_s$, where m, n are integers and $m \leq q$.

We propose two-dimensional information entropy as a measure to quantify the randomness of Δs_i and Δt_i. The two-dimensional information entropy is actually made up of two marginal entropies: $H(\Delta s_i)$ and $H(\Delta t_i)$, assuming independence of the random variables Δs_i and Δt_i.

Firstly, we calculate histograms of discrete random variables Δs_i and Δt_i within certain time interval, which is called frame duration and denoted with t_0. Secondly, we calculate information entropies $H(\Delta s_i)$ and $H(\Delta t_i)$

$$H(\Delta s_i) = \sum_{i=1}^{I} P(\Delta s_i) ld \frac{1}{P(\Delta s_i)}\,, \tag{1}$$

$$H(\Delta t_i) = \sum_{i=1}^{I} P(\Delta t_i) ld \frac{1}{P(\Delta t_i)}\,, \tag{2}$$

where I denotes number of intervals Δt_i within a frame, *ie*

$$t_0 = \sum_{i=1}^{I} \Delta t_i\,. \tag{3}$$

It will be shown that the proposed two-dimensional information entropy is useful feature domain which represents speaker specific characteristic suitable for text-independent speaker recognition.

Description of experimental testbed

Experimental testbed consists of voice activity detector, A/D converter and two-dimensional information entropy extractor, as shown in Fig. 2.

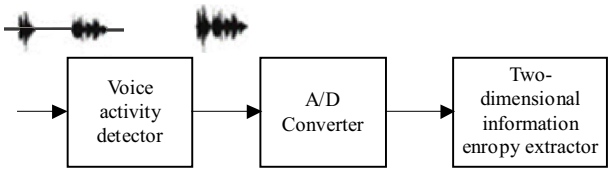

Fig. 2. Experimental testbed

The function of the voice activity detector is to extract speech segments from the speech signal. A simple method [26], based on two audio features (signal energy and spectral centroid), for extraction of speech segments by removing the silence is used in testbed.

Once the speech segments have been extracted, speech signal is sampled at $f_s = 1/T_s = 8$ kHz sampling rate with an 8-bits A/D convertor, *ie* each sample is quantized into one of $q = 256$ levels.

The most important step in the speaker recognition process is to extract features from the analyzed signal. In two-dimensional information entropy extractor, speech signal is windowed into frames and processed sequentially. Calculations are performed according to relations (1) and (2).

3 NUMERICAL RESULTS

Speech signal database is formed of six the most frequent speakers from Serbian parliament, three of them are males (denoted with M1, M2 and M3), while the remaining three are females (denoted with F1, F2 and F3). Duration of speech signal of any of them is shortly below 4 min.

Histograms of discrete random variables Δs_i and Δt_i are calculated for all six speakers. Using these histograms, information entropies $H(\Delta s_i)$ and $H(\Delta t_i)$ are calculated according to relations (1) and (2). $H(\Delta s_i)$ and $H(\Delta t_i)$ represents coordinates in two-dimensional information entropy domain. Different frame durations for calculating histograms and information entropies $H(\Delta s_i)$ and $H(\Delta t_i)$ are considered: $t_0 = 10$ s, 20 s and 30 s.

Obtained results for each specific frame can be represented by point which is defined by ordered pair $H(\Delta s_i), H(\Delta t_i)$ in two-dimensional information entropy domain. Numerical results in $H(\Delta s_i), H(\Delta t_i)$ plane are presented in Fig. 3, subfigures (a), (b) and (c), for $t_0 = 10$ s, 20 s and 30 s, respectively. From this figure one can conclude that two-dimensional information entropy points, obtained for one speaker, are clustered. In other words, within-speaker variability from frame to frame is significantly smaller relative to between-speaker variability. Following the terminology from vector quantization (VQ) based approach [27, 28], ordered pair $H(\Delta s_i), H(\Delta t_i)$ is called speaker's feature vector and the speaker's model is formed by clustering the speaker's feature vectors. In VQ-based approach, the speakers' models are formed by clustering the K speakers' feature vectors in K non-overlapping clusters.

Coordinates of the centre of the cluster are calculated as

$$\overline{H(\Delta s_i)} = \sum_{i=1}^{N} H(\Delta s_i)\,, \tag{4}$$

$$\overline{H(\Delta t_i)} = \sum_{i=1}^{N} H(\Delta t_i)\,, \tag{5}$$

where N denotes number of the points in the cluster. For $t_0 = 10$ s is $N \cong 20$, for $t_0 = 20$ s is $N \cong 10$, while for $t_0 = 30$ s is $N \cong 6$. According to terminology from [27, 28], each cluster is represented by a code vector which is the centroid (average vector) of the cluster. VQ model, also known as centroid model, is one of the simplest text-independent speaker models.

Radius of each cluster, presented in Fig. 3, is calculated as standard deviation of distances between points and the centre of the cluster

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\{[H(\Delta s_i), H(\Delta t_i)] - [\overline{H(\Delta s_i)}, \overline{H(\Delta t_i)}]\}^2} \tag{6}$$

From Fig. 3(a), obtained for frame duration 10 s, one can see that clusters are overlapping. Gaussian Mixture Model (GMM) can be considered as an extension of the VQ model, in which the clusters are overlapping [29]. GMM is composed of a finite mixture of multivariate Gaussian components. Hence, a feature vector is not assigned to the nearest cluster as in VQ model, but it has a nonzero probability of originating from each cluster.

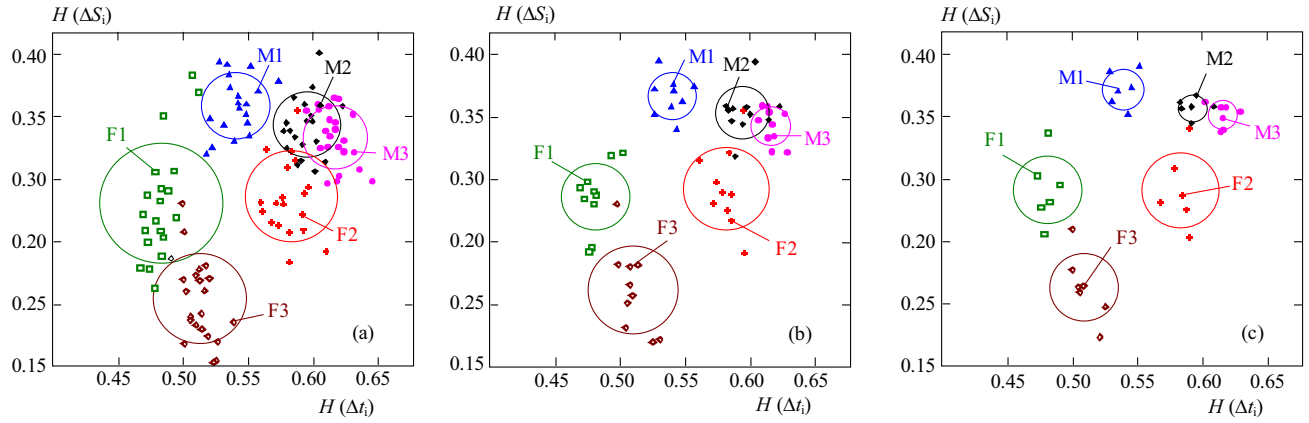

Fig. 3. Two-dimensional information entropy for six speakers, males are denoted with M1, M2 and M3, females are denoted with F1, F2 and F3: (a) — Frame duration 10 s, (b) — Frame duration 20 s, (c) — Frame duration 30 s

From Fig. 3, one can see that standard deviation of two-dimensional information entropy of a speaker is reduced as the frame duration is increased. In addition, standard deviation of two-dimensional information entropy is higher for females than for males.

Although the frame duration of 10–30 s seems to be long, it is comparable with actual systems. Recently, it was announced that Barclays Wealth was to use speaker recognition to verify the identity of telephone customers within 30 seconds of normal conversation [30].

4 CONCLUSION

Two-dimensional information entropy is useful feature domain for text-independent speaker recognition. Although the validation is performed using small dataset, obtained results clearly show that this feature can be used to discriminate between speakers. Two-dimensional information entropy is very accurate in gender identification. The most significant factor affecting automatic speaker recognition performance is variability of signal characteristics from trial to trial, *ie* between-trial variability. Variations arise from the speaker him/herself, from differences in recording and transmission conditions, and from different noise environment. These topics are subject of further researches.

References

[1] Biometric Recognition: Challenges and Opportunities (Pato, J. , Millett, L. I., eds.), National Academies Press, Washington, 2010.

[2] TOGNERI, R.—PULLELLA, D.: An Overview of Speaker Identification: Accuracy and Robustness Issues, IEEE Circuits and Systems Magazine, Second quarter (2011), 23–61.

[3] CAMPBELL, J. P.: Speaker Recognition: A Tutorial, Proc. of the IEEE **85** No. 9 (1997), 1437–1462.

[4] ROSE, P.: Forensic Speaker Identification, Taylor & Francis, London, 2002.

[5] KINNUNEN, T.: Spectral Features for Automatic Text-Independent Speaker Recognition, Licentiate's thesis, University of Joensuu, Joensuu, Finland, 2003.

[6] KINNUNEN, T.: Optimizing Spectral Feature Based Text-Independent Speaker Recognition, PhD thesis, University of Joensuu, Joensuu, Finland, 2005.

[7] JIN, Q.—SCHULTZ, T.—WAIBEL, A.: Phonetic Speaker Identification, Proc. of the Int. Conference of Spoken Language Processing (ICSLP 2002), Denver, CO, Sep 2002, pp. 1345-1348.

[8] BACHOROWSKI, J. A.—OWREN, M. J.: Acoustic Correlates of Talker Sex and Individual Talker Identity are Present in a Short Vowel Segment Produced in Running Speech, Journal of Acoust. Soc. America. **106** No. 2 (1999), 1054–1063.

[9] ADAMI, A. G.: Modeling Prosodic Differences for Speaker Recognition, Speech Communication **49** No. 4 (2007), 277–291.

[10] FURUI, S.: Digital Speech Processing, Synthesis, and Recognition, 2nd ed., Marcel Dekker, New York, 2001.

[11] SIVAKUMARAN, P.—ARIYAEEINIA, A.—LOOMES, M.: Sub-Band Based Text-Dependent Speaker Verification, Speech Communication **41** No. 2-3 (2003), 485–509.

[12] DAVIS, S.—MERMELSTEIN, P.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, IEEE Trans. Acoustics, Speech, Signal Processing **28** No. 4 (1980), 357–366.

[13] HERMANSKY, H.: Perceptual Linear Predictive (PLP) Analysis of Speech, Journal Acoust. Soc. America, **87** No. 4 (1990), 1738–1752.

[14] MAMMONE, R.—ZHANG, X.—RAMACHANDRAN, R.: Robust Speaker Recognition: a Feature Based Approach, IEEE Signal Processing Magazine **13** No. 5 (1996), 58–71.

[15] NOLAN, F.: The Phonetic Bases of Speaker Recognition, Cambridge, 1983.

[16] ANTAL, M.: Phonetic Speaker Recognition, Proc. of 7th Int. Conference Communications, Bucharest, Romania, June 2008, pp. 67–72.

[17] DEHAK, N.—KENNY, P.—DUMOUCHEL, P.: Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification, IEEE Trans. Audio, Speech and Language Processing **15** No. 7 (2007), 2095–2103.

[18] MILIVOJEVIĆ, Z. N.—BRODIĆ, D.: Estimation of the Fundamental Frequency of the Speech Signal Compressed by G.723.1 Algorithm Applying PCC Interpolation, Journal of Electrical Engineering, **62** No. 4 (2011), 181–189.

[19] KINNUNEN, T.—LI, H.: An Overview of Text-Independent Speaker Recognition: From features to Supervectors, Speech Communication **52** No. 1 (2010), 12–40.

[20] SEDLAK, V.—DURACKOVA, D.—ZALUSKY.—KOVACIK, T.: Intelligibility Assessment of Ideal Binary-Masked Noisy Speech with Acceptance of Room Acoustic, Journal of Electrical Engineering **65** No. 6 (2014), 325–332.

[21] PRIBIL, J.—PRIBILOVA, A.—DURACKOVA, D.: An Experiment with Spectral Analysis of Emotional Speech Affected by Orthodontic Appliances, Journal of Electrical Engineering **63** No. 5 (2012), 296–302.

[22] O'SHAUGHNESSY, D. Automatic Speech Recognition: History, Methods and Challenges: Pattern Recognition **41** (2008), 2965–2979.

[23] VASEGHI, S. V.: Multimedia Signal Processing: Theory and Applications in Speech, Music and Communications, John Wiley & Sons, 2007.

[24] BRUMMER, N.—du PREEZ, J.: Application Independent Evaluation of Speaker Detection, Computer Speech and Language **20** No. 2-3 (2006), 230–275.

[25] ARONOWITZ, H.—BURSHTEIN, D.: Efficient Speaker Recognition using Approximated Cross Entropy (ACE), IEEE Transactions on Audio, Speech and Language Processing **15** No. 7 (Sep 2007), 2033–2043.

[26] GIANNAKOPOULOS, T.: Silence Removal in Speech Signals, MATLAB Central, March 2014, available from: http://www.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-signals, accessed on December 14, 2014.

[27] SOONG, F. K.—ROSENBERG, A. E.—JUANG, B. H.—RABINER, L. R.: A Vector Quantization Approach to Speaker Recognition, AT&T Technical Journal **66** No. 2 (Mar-Apr 1987), 14–26.

[28] LINDE, Y.—BUZO, A.—GRAY, R.: An Algorithm for Vector Quantizer Design, IEEE Trans. on Communications **28** No. 1 (1980), 84–95.

[29] REYNOLDS, D. A.—ROSE, R. C.: Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models, IEEE Trans. on Speech and Audio Processing **3** No. 1 (1995), 72–83.

[30] Barclays International Banking, available from: https://wealth.barclays.com/en_gb/internationalwealth/manage-your-money/banking-on-the-power-of-speech.html, accessed on December 14, 2014.

Received 26 February 2015

Boško Božilović was born in Belgrade, Serbia, in 1978. He received Dipl Eng and MSc degrees from the Faculty of Electrical Engineering, University of Belgrade, in 2003, and 2012, respectively. He is a Director of ICT at VLATACOM, R&D Center. His research interests are in the areas of biometrics, forensics and digital security. He has authored or co-authored several peer-reviewed journal and conference papers and holds one patent. Currently he is working towards his PhD degree.

Branislav M. Todorović was born in Belgrade, Serbia, in 1959. He received Dipl Eng and MSc degrees from the Faculty of Electrical Engineering, University of Belgrade, and PhD degree from the Faculty of Technical Sciences, University of Novi Sad, in 1983, 1988 and 1997, respectively. He is a Senior Research Fellow at the RT-RK, Institute for Computer Based Systems, and a Full Professor at the Military Academy, University of Defence, Belgrade. He is also with VLATACOM, R&D Center, Belgrade. Prior to joining RT-RK, he was with the Institute of Microwave Techniques and Electronics IMTEL-Komunikacije, Centre for Multidisciplinary Research, and the Military Technical Institute (VTI, Institute of Electrical Engineering) in Belgrade. His research interests are in the wide area of radio telecommunications and digital signal processing. He has authored or co-authored more than 100 peer-reviewed journal and conference papers and three books.

Miroslav Obradović was born in Belgrade, Serbia, in 1978. He is a Senior Software Developer at VLATACOM, R&D Center, Belgrade.