INFORMATICA, 1993, Vol.4, No.1–2, 45–56

SPEAKER IDENTIFICATION

Antanas LIPEIKA and Joana LIPEIKIENĖ

Institute of Mathematics and Informatics
2600 Vilnius, Akademijos St.4, Lithuania

Abstract. Speaker identification problem is investigated. The identification is carried out comparing feature vectors (parameters of LPC model) of the "criminal" and "suspicious" speakers. Both likelihood ratio and cepstral distances are used for comparing feature vectors. The feature vectors are extracted from pseudostationary parts of speech utterances. The identification approach is suitable for text-dependent and text-independent identification. Experimental results illustrate the performance of the algorithm.

Key words: likelihood ratio distance, cepstral distance, pseudostationary segments, distance measure.

1. Introduction. The automatic speaker identification problem [1] is very urgent in the forensic examination. It is more difficult to identify a speaker by his speech phonogram than, for example, by finger-prints [2]. The latters are unique and their picture in practice does not change all the life, meanwhile human voice changes in time, it depends on the emotional state and other factors. Besides, a voice phonogram is distorted when recording (influence of an environment noise, imperfect recording equipment, etc.). Therefore this investigation field is being intensively developed [3 – 12]. The problems of selection of features, a structure of an identification system and a decision rule are still urgent.

Possibilities of speaker identification by pseudostationary segments are investigated in this paper. When pronouncing a speech utterance, a vocal tract is sometimes fixed for a short period, therefore there occurs a possibility to measure parameters of vocal tract and to identify a speaker using phonograms.

For detection of pseudostationary segments we applied the

method [13], which is often used in the speech recognition. According to this method two neighbouring frames are compared, calculating the likelihood ratio distance between them. When the distance exceeds the threshold, chosen in advance, it is warned about the end of a pseudostationary segment. Coefficients of the linear prediction model (LPC), estimated by the correlation method [15], are used as feature parameters. The speaker identification is carried out comparing the average distances between "criminal" and "suspicious" speakers.

2. Statement of the problem. Consider the typical situation, which occurs in solving the speaker identification problem. Let us have phonograms of the true speaker X (criminal) and of n "suspicious" speakers $A_1, A_2, \ldots, A_n$ and it is possible to detect pseudostationary segments in these phonograms. When solving the problem we must answer the following question:

Which of the "suspicious" speakers is closest to the "criminal X*?*

3. Detection of pseudostationary segments. The phonogram considered is divided into frames which are moved with respect to one another by M points (a step of a frame is M). A spectral preemphasis of a signal y_t for all frames is done with the filter $x_t = y_t - 0{,}94y_{t-1}$ [16], then each frame is weighted by the Hamming window [17]. After that, according to the Durbin algorithm [18] LPC parameters are estimated and autocorrelation coefficients are calculated. Further, using autocorrelation coefficients of the LPC model of a previous frame and a correlation function of a next frame, divided by a square LPC model amplification coefficient, estimated in this frame, we calculate the likelihood ratio distance [14] for all neighbouring pairs of frames. If a distance between two neighbouring frames is less than a preassigned threshold (the threshold is chosen experimentally), we draw a conclusion that moving by a frame step does not change a spectral structure of a signal. Since we are not interested in very short pseudostationary segments, we compare them with the threshold of the minimal

pseudostationary interval and leave for further investigation only those segments which are longer than this threshold.

The likelihood ratio distance has the spectral interpretation [14]:

$$d_{LR}(1/\widetilde{A}, 1/A) = \int_{-\pi}^{\pi} \frac{|A(e^{j\theta})|^2}{|\widetilde{A}(e^{j\theta})|^2} \frac{d\theta}{2\pi} - 1, \tag{1}$$

where $1/A$ is the LPC model transfer function of the first frame, $1/\widetilde{A}$ is the LPC model transfer function of LPC model of the second frame. As the spectral densities of LPC model of the first and second frames are

$$S(\theta) = 2b^2/|A(e^{j\theta})|^2, \qquad \widetilde{S}(\theta) = 2\widetilde{b}^2/|\widetilde{A}(e^{j\theta})|^2$$

we may interpret $1/|A(e^{j\theta})|^2$ and $1/|\widetilde{A}(e^{j\theta})|^2$ as corresponding spectral densities of LPC model of the first and second frame with an amplification equal to 1. Then the likelihood ratio distance may be expressed by spectral densities as follows:

$$d_{LR}(\widetilde{S}, S) = \int_{-\pi}^{\pi} \frac{\widetilde{S}(\theta)\widetilde{b}^2}{S(\theta)/b^2} \frac{d\theta}{2\pi} - 1 \,. \tag{2}$$

Due to a great computation amount it is not convenient to use expression (2), the likelihood ratio distance is usually calculated in the time domain:

$$d_{LR}(\widetilde{S}, S) = \left\{ \frac{r_x(0)}{\widetilde{b}^2} r_a(0) + 2 \sum_{i=1}^{p} \frac{r_x(i)}{\widetilde{b}^2} r_a(i) \right\} - 1 \,, \tag{3}$$

where $r_x(i)$ is the autocorrelation function of a signal in the second frame, $r_a(i)$ are autocorrelations of parameters of LPC model for the first frame:

$$r_a(i) = \sum_{k=0}^{p-i} a_{k+i} a_k, \quad i = 0, 1, 2, \ldots, p \,, \tag{4}$$

p is the order of LPC model, $\widetilde{b}$ is the amplification coefficient of LPC model for the second frame.

4. Detection of the "suspicious" closest to the true speaker X. If there are not one but several "suspicious" speakers there is need to detect which of them is "closest" to the speaker X. Let us have N_X pseudostationary segments of the speaker X and N_{A_k} $(k = 1,2,\ldots,n)$ pseudostationary segments of n suspicious. Let us determine all possible distances $d_{ji}(X, A_k)$ between the speaker X and suspicious A_k, $k = 1,2,\ldots,n$. Since pseudostationary intervals of the speaker X and of the suspicious form $n+1$ cluster in a multivariate space of features, let us estimate the distance between clusters corresponding to X and A_k:

$$D_{XA_k} = \frac{1}{N_X}\sum_{j\in X}\min_{i\in A_k} d_{ji}(X, A_k) + \frac{1}{N_{A_k}}\sum_{i\in A_k}\min_{i\in X} d_{ji}(X, A_k). \quad (5)$$

The closest suspicious is detected as

$$\hat{I} = \arg\min_{1\leqslant k\leqslant n} D_{XA_k} . \quad (6)$$

The expressions (5) and (6) can be used for detection of the closest suspicious in the text-dependent and text-independent identification.

5. Choice of a distance measure. When detecting the pseudostationary segments we have used the likelihood ratio distance (3). But one can see from formula (1) that this measure is not symmetric, i.e.

$$d_{LR}(\tilde{S}, S) \neq d_{LR}(S, \tilde{S}) . \quad (7)$$

It is not shortcoming in the detection of pseudostationary segments because a threshold is not high, meanwhile assymetry appears when values of distance are large. But when calculating the distances between speakers (clusters) distances between speech frames may be large and assymetry is undesirable. So we used the symmetric distance [21]

$$d(\tilde{S}, S) = \frac{d_{LR}(\tilde{S}, S) + d_{LR}(S, \tilde{S})}{2} \quad (8)$$

for calculating the distortions.

When solving speaker identification problem the main point is a choice of a feature system and of a distance measure. Apart from the likelihood ratio distance other distance measures are used, too. The cepstral distance [3 – 5],[21] and its various modifications are mainly widespread. According to [22 – 23], the cepstral coefficients $c_1, \ldots, c_L$ can be calculated from LPC coefficients as follows

$$
\begin{aligned}
c_0 &= \ln b^2, \\
c_1 &= -a_1, \\
c_n &= -\sum_{k=1}^{n-1}(1 - k/n)a_k c_{n-k} - a_n, \quad n = 2, \ldots, p, \qquad (9) \\
c_n &= -\sum_{k=1}^{n-1}(1 - k/n)a_k c_{n-k}, \quad n = p+1, \ldots, L. \qquad (10)
\end{aligned}
$$

The cepstral distance measure between two frames with corresponding LPC coefficients $(a_1, \ldots, a_p, b)$, $(\tilde{a}_1, \ldots, \tilde{a}_p, \tilde{b})$ may be defined as

$$
d_{\text{cep}}(L) = [u(L)]^2 = (c_0 - \tilde{c}_0)^2 + 2\sum_{k=1}^{L}(c_k - \tilde{c}_k)^2. \qquad (11)
$$

It is important to know [24] that, as L increases, $u(L)$ approaches d_2 from below and

$$
\lim_{L\to\infty} u(L) = d_2 \quad , \qquad (12)
$$

where d_2 is root mean square log spectral measure (L_2 norm) and has the following spectral interpretation

$$
d_2^2 = \int_{-\pi}^{\pi} \left| \ln \frac{\tilde{S}(\theta)}{S(\theta)} \right|^2 \frac{d\theta}{2\pi} \quad , \qquad (13)
$$

where $S(\theta)$, $\tilde{S}(\theta)$ are spectral densities corresponding to the first and second frame respectively.

Usually we wish the distance to be gain independent and assume the amplification equal to 1 (i.e., $b = \tilde{b} = 1$). Then we can

rewrite (13) in the form

$$d_2^2 = \int_{-\pi}^{\pi} \left| \ln \frac{\widetilde{S}(\theta/\widetilde{b}^2}{S(\theta)/b^2} \right|^2 \frac{d\theta}{2\pi} - \int_{-\pi}^{\pi} \left| \ln \frac{|A(e^{j\theta})|^2}{|\widetilde{A}(e^{j\theta})|^2} \right| \frac{d\theta}{2\pi} \quad . \tag{14}$$

Cepstral distance satisfies the following properties

1) $d(x,y) = d(y,x)$ symmetry;
2) $d(x,y) > 0$ for $x \neq y$ and $d(x,x) = 0$ positive definitness;
3) $d(x,y)$ has a physically meaningful interpretation in the frequency domain;
4) It can be efficiently evaluated.

6. Experiments. For experiments we have designed software in C language for the computer IBM PC/AT. Phonograms of the speech signal analysed were digitized in a 12 bits A/C converter at a rate of 10000 Hz. When computing the following parameters were taken:

- frame length – 250 signal samples(25 ms)
- frame step – 50 signal samples(5 ms)
- threshold for detecting of pseudostationary segments – 0.07
- threshold for a pseudostationary segment length – 250 samples (25ms).
- order of LPC model – 10.

EXAMPLE A.

Phonograms of six speakers (five men and a woman) were recorded. Every speaker in two sessions repeated a Lithuanian word "langas" for ten times. A phonogram of each investigated speaker recorded in each session in turn was regarded as a phonogram of "criminal" and the closest "suspicious" was determined. Pseudostationary intervals in every word were detected automatically.These segments were divided into equal (25 ms) frames. There were from 47 to 131 such frames for each investigated speaker in each session (the number of frames is indicated in Table 1).When detecting the closest "suspicious" 91.6% of true answers were obtained. The experimental results are given in Table 1.

Table 1. Text-dependent identification, Lithuanian word "langas"

Investigated speakers	Session number	Number of repetitions of a word	Overall number of frames obtained from stat. intervals	Results of the identification (true answer – yes, wrong answer – no) likelihood ratio distance
M_1	I	10	123	yes
	II	10	114	yes
M_2	I	10	130	yes
	II	10	128	yes
M_3	I	10	100	yes
	II	10	131	yes
M_4	I	10	47	yes
	II	10	112	yes
M_5	I	10	121	yes
	II	10	105	yes
M_6	I	10	116	yes
	II	10	97	no
		positive results:		91.6%

EXAMPLE B.

In further experiments we used text independent utterances. Their duration was from 3 to 5 minutes. We divided these utterances into two equal parts. Pseudostationary segments of voiced sounds were selected and used for identification. Experiments were fulfilled using the likelihood ratio distance and the cepstral distance.

In Table 2 the results of identification are presented for 10 men. Both,cepstral and likelihood ratio distances gave the same results –90% of true answers.

Table 3 illustrates similar results for five women. Both distances gave 100% of true answers.

In Table 4 results of identification using telephone speech are given. For recording the telephone speech a standard telephone line was used. In the experiment the same speech material was used

Table 2. Text independent identification. Voiced sounds. 10 men

Investi-gated speakers	Session number	Overall number of frames obtained from stat. intervals	Results of the identification (true answer – yes, wrong answer – no)	
			cepstral distance	likelihood ratio distance
M_1	I	73	yes	yes
	II	106	yes	yes
M_2	I	108	yes	yes
	II	98	yes	yes
M_3	I	126	yes	yes
	II	131	yes	yes
M_4	I	123	yes	yes
	II	116	yes	yes
M_5	I	107	yes	yes
	II	119	yes	yes
M_6	I	146	yes	yes
	II	148	yes	yes
M_7	I	86	no	no
	II	72	no	no
M_8	I	62	yes	yes
	II	65	yes	yes
M_9	I	106	yes	yes
	II	98	yes	yes
M_{10}	I	99	yes	yes
	II	88	yes	yes
positive results:			90%	90%

as in the previous experiment and speech utterances for four men and one woman were recorded. The identification results for both distances were 100%.

7. Conclusions. The main difference between this research and known publications is that we have used pseudostationary segments of speech utterances for speaker identification. Identification was fulfilled by using the likelihood ratio distance and the cepstral distance between speech frames.

Table 3. Text independent identification. Voiced sounds. 5 women

Investigated speakers	Session number	Overall number of frames obtained from stat. intervals	Results of the identification (true answer – yes, wrong answer – no)	
			cepstral distance	likelihood ratio distance
W_1	I	106	yes	yes
	II	121	yes	yes
W_2	I	114	yes	yes
	II	95	yes	yes
W_3	I	100	yes	yes
	II	84	yes	yes
W_4	I	88	yes	yes
	II	78	yes	yes
W_5	I	79	yes	yes
	II	85	yes	yes
	positive results:		100%	100%

Table 4. Text independent identification. Voice sounds. 4 men and a woman. Telephone speech

Investigated speakers	Session number	Overall number of frames obtained from stat. intervals	Results of the identification (true answer – yes, wrong answer – no)	
			cepstral distance	likelihood ratio distance
M_1	I	123	yes	yes
	II	108	yes	yes
M_2	I	142	yes	yes
	II	124	yes	yes
M_3	I	93	yes	yes
	II	84	yes	yes
M_4	I	97	yes	yes
	II	74	yes	yes
W_5	I	86	yes	yes
	II	90	yes	yes
	positive results:		100%	100%

We have investigated which distance measure (likelihood ratio or cepstral) provides a higher identification accuracy. The results of experiments showed that there was no significant difference between the likelihood ratio and the cepstral distance measures. It seems to be a little greater "reliability reserve" when the cepstral distance is used for comparison of two frames of speech.

8. Acknowledgment. The author wishes to thank B. Šalna, V. Malinauskas and Ž. Apanavičiūtė of the Department of Phonoscopic Expertise, the Lithuanian Institute of Forensic Expertise for valuable discussions, their help in obtaining the database and support in performing experiments.

REFERENCES

[1] Ramishvili, G. (1984). Automatic speaker recognition by voice. *Radio i sviaz.* Moscow (in Russian).

[2] Ramishvili, G., G. Chikoidze (1991). Forensic investigation of speech phonograms and speaker identification. *MECNIEREBA*. Tbilisi (in Russian).

[3] Noda, H. (1989). On the use of the information on individual speakers position in the parameter space for speaker recognition. *Proc. of the ICASSP*, 516–519.

[4] Xu, L., J. Oglesby, and J. Mason (1989). The optimization of perceptually based features for speaker identification. *Proc. of the ICASSP*, 520–523.

[5] Naik, J., L. Netsch, and G. Doddington (1989). Speaker verification over long distance telephone lines. *Proc. of the ICASSP*, 524–527.

[6] Noda, H. (1988). Frequency-warped spectral distance measures for speaker verification in noise. *Proc. of the ICASSP*, 576–579.

[7] Zheng, Y., and B. Yuan (1988). Text-dependent speaker identification using circullar hidden Markov models. *Proc. of the ICASSP*, 580–582.

[8] Velius, N. (1988). Variants of cepstrum based speaker identity verification. *Proc. of the ICASSP*, 583–586.

[9] Nakasone, N., and C. Nelvin (1988). Computer assisted voice identification system. *Proc. of the ICASSP*, 587–590.

[10] Wilbur, J., and F. Taylor (1988). Consistent speaker identification via Wig-

ner smoothing techniques. *Proc. of the ICASSP*, 591–594.

[11] Li, K. and J. Porter (1988). Normalizations and selection of speech segments for speaker recognition scoring. *Proc. of the ICASSP*, 595–596.

[12] Attili, J. and M. Savic (1988). A TMS32020 -based real time text-independent, automatic speaker verification system. 599–602.

[13] Lowerre, B. The Harpy speech understanding system. In Wayne A. Lea (Ed.), *Prentice Hall.*

[14] Juang., D. Wong, and A. Gray (1982). Distortion performance of vector quantization for LPC voice coding. *IEEE Trans. on Acoust. Speech and Signal Processing*, **ASSP-30**(2), 294–304.

[15] Markel, J., and A.Jr. Gray (1976). *Linear Prediction of Speech.* Springer - Verlag, Berlin, Heidelberg, New York.

[16] Tribolet, J., L. Rabiner, and M. Sondhi (1979). Statistical properties of an LPC distance measure. *IEEE Trans. on Acoust. Speech and Signal Processing*, **ASSP-27**(5), 550–558.

[17] Marfle, S.Jr. (1987). *Digital Spectral Analysis with applications.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[18] Rabiner, L., and R. Schafer (1978). *Digital Processing of Speech Signals.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[19] Box, G., and G. Jenkins (1970). *Time Series Analysis forecasting and control.* HOLDEN-DAY, San Franc.

[20] Mandel, I. (1988). Cluster analysis. *Finansy i statistika.* Moscow (in Russian).

[21] Bastura, C., W. Majewski, W. Myslecki and J. Zalewski (1990). *Speech signal parametrization methods for automatic speaker recognition.* Wydawnictwo Politechnili Wroclawskiej, Wroclaw.

[22] Atal, B. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Acoust. Soc. Amer.*, **55**(6), 1304–1312.

[23] Atal, B. (1976). Automatic recognition of speakers from their voices. *Proc. IEEE*, **64**(4), 460–475.

[24] Gray A., and J. Markel (1976). . *Distance Measures for Speech Processing*, **ASSP-24**(5), 380–391.

Received January 1993

A. Lipeika is a Candidate of Technical Sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics. Scientific interests include: processing and recognition of random processes, detection of changes in the properties of random processes.

J. Lipeikienė is a Candidate of Technical Sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics. Scientific interests include: processing of random signals, robust methods for determination of change-points in the properties of random processes, data compression.