COMPARATIVE ANALYSYS OF SPEECH PARAMETERS FOR THE DESIGN OF SPEAKER VERIFICATION SYSTEMS

A. F. Souza[1], *Student Member*, *IEEE*, M. N. Souza[1,2], *Member*, *IEEE*
[1] Biomedical Engineering Program – COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
[2]Electronic Department E.E., Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.

Abstract – **Speaker verification systems are basically composed of three stages: feature extraction, feature processing and comparison of the modified features from speaker voice and from the voice that should be verified. Many features have been used in the first stage, although the current literature has not already shown the best of them. Based on the biometrics hypothesis, which states that each individual has a physical characteristic that distinguishes itself from the others, this paper realized a comparison between 12 classical widely used parameters, in order to investigate the biometrics hypothesis. The obtained results point out those parameters directly correlated to speaker's anatomy which are among the best ones that can be used in the development of speaker verification systems.**

Keywords: speaker verification, biometrics, area function.

I. INTRODUCTION

Physical characteristics of the subjects have been used in several scientific works concerning to subject verification, and its use ranges from security to forensic applications [1]. The use of such characteristics qualifies the so-called biometrics technique, which states that each subject exhibits some individual patterns that distinguish he/she among others. It has been reported that this approach can present some advantages when compared with others classical features, and that it is generally more reliable and secure [2].

The aim of most of speaker recognition/verification studies is to develop a free-time system, not biased, fast, free-text, that present the same accuracy of human being in speaker recognition. The motivation for such studies is to generate more robust and reliable systems that can be used in financial security, psychological evaluation [3], vocal tract evaluation [4], as well as to be accept in forensic area.

Speaker verification systems use many methods, such as neural networks [5, 6, 7], gaussian mixture model [8, 9], data fusion [10], prony technique [11], cohort models [12], orthogonal linear prediction [13], among others, to perform the comparison/classification of features, that belong normally to the set LPC, PARCOR, AR, Area Function, Log Area, etc..

The aim of this paper is to perform an objective comparison between a set of ordinarily used speech-derived parameters to verify if the biometrics hypothesis holds and if the parameters directly correlated with the anatomic aspects (Fig. 1) of the speaker really present a good performance in the design of a speaker recognition/verification system. The reason to do such work is the contests presented in the literature concerning to the accuracy and robustness of several parameters. Imperl [11] and Kishore [5], for example, used cepstral parameters, Furui [16] used log area ratios, Sambur [13] used LPC parameters and all of them got good results. For this reason, and considering the beginning of a speaker verification design, it is reasonable to investigate the set of the best speech-dependent features.

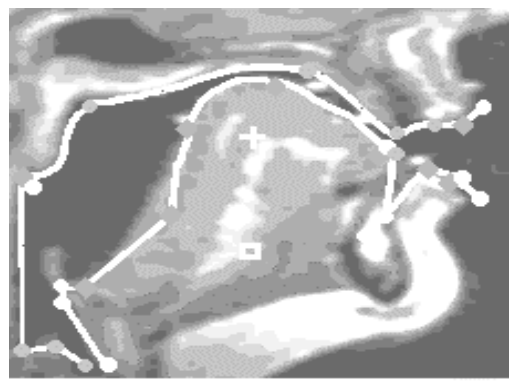

Fig. 1. Magnetic resonance image of the vocal tract.

It will be shown the results of the comparison among twelve classically used parameters, using the same preprocessing and comparison stages in the rest of the system, the Sambur's technique [13]. Such results confirm the hypothesis that parameters linked to biometrics approach are among the best studied features, suggesting their adoption in the design of speaker verification systems aimed to be widely used, including forensic applications [1].

II. METHODOLOGY

A. Signal processing

Since the system designed by Sambur [13] will be used as the fixed part in the comparison, it is important to summarize such approach.

The several utterances ($l = 1, 2, \ldots, L$) of the speech signal from the m-th speaker that will compound the base to the verification system are initially divided in J_{lm} fixed-size frames and stored in L matrices, where each column corresponds to a signal frame. From such matrices p coefficients are extracted from each column, using the desired

This work has been partially supported by the Brazilian research agencies Capes, Faperj and Pronex.

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Comparative Analysys of Speech Parameters for the Design of Speaker Verification Systems	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Biomedical Engineering Program COPPE, Federal University of Rio de Janeiro Brazil	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.	
Abstract	
Subject Terms	
Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU
Number of Pages 4	

parameters extraction technique. This procedure results in new L matrices A_{lm}.

From the parameters matrices A_{lm}, covariance matrices R_{lm} are calculated. For all the m speakers the system will be trained to verify the so-called reference covariance matrix ($R_{ref}^{(m)}$), which is calculated by the weighted average of the R_{lm} matrices, being the weights the number of frames in the respective matrices (1). This procedure reduces the random estimation error [13].

$$R_{ref}^{(m)} = \frac{1}{\sum_{l=1}^{L} J_{lm}} \sum_{l=1}^{L} J_{lm} R_{lm} \tag{1}$$

where J_{lm} is the number of frames in the l-th utterance for the m-th speaker.

Given the reference covariance matrix, the statistical variance (eigenvalue) of each orthogonal parameter is first found by solving the set of simultaneous equations

$$\left| R_{ref}^{(m)} - \lambda_{ref} I \right| = 0 . \tag{2}$$

The mutually orthogonal eigenvectors (b_i) are then derived as solution of the equation

$$\lambda_{iref} b_i = R_{ref}^{(m)} b_i \quad i = 1, 2, \ldots, p . \tag{3}$$

The eigenvectors (b_i) associated to the reference covariance matrix lead to a conversion matrix. The i-th orthogonal parameter f_i in the j-th frame of the A_{lm} matrix, for the m-th speaker, is obtained from the product of the conversion matrix and the parameter matrix, as demonstrated in (4). The average value of the i-th orthogonal parameter for the m-th speaker is given by (5).

$$f_{ijm} = A_{jlm}^{|} \cdot b_i \tag{4}$$

$$\overline{f_{im}} = \frac{1}{\sum_{l=1}^{L} J_{lm}} \sum_{l=1}^{L} \sum_{j=1}^{J_{lm}} f_{ijm} \tag{5}$$

In the verification process the unknown speaker's voice will be initially processed in the same way. This process is based on dissimilarity between the unknown speaker's orthogonal parameter set and the analogous set for each of the m speakers the system is able to verify. Such dissimilarity is based on the distance between the two sets of parameters, calculated as

$$d_m = \sum_{i=p}^{p} \left(\frac{\overline{f}_{im} - Z_i}{\sqrt{\lambda_{im}}} \right)^2 \overline{J}_m \tag{6}$$

where Z_i is the mean value of the i-th orthogonal parameter calculated across the utterance of the unknown speaker by b_{im} (3); λ_{im} is the reference eigenvalue for the i-th orthogonal parameter of the m-th speaker; $(p' - 1)$ is the number of the first most important orthogonal parameters not included in the summation, and $\overline{J_m}$ is the average number of frames in the utterance of the m-th speaker's design set, calculated as

$$\overline{J_m} = \frac{1}{L} \sum_{l=1}^{L} J_{lm} \tag{7}$$

Ordinarily speaker verification process adopt a threshold decision, below which the unknown speaker is claimed the speaker being verified. This process is illustrated in Fig. 2.

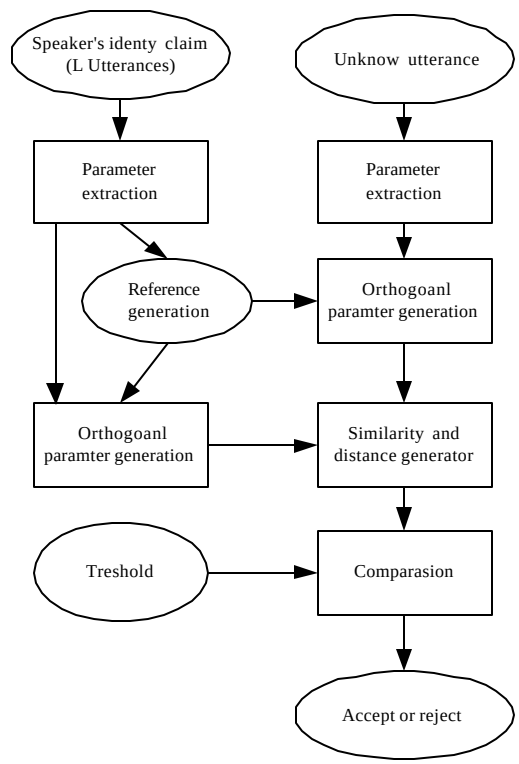

Fig. 2. Schematic of a speaker's verification generic process using the Sambur's method [13].

The above version of speaker verification system was used in the comparison of the twelve speech-dependent studied parameters, that was the only part that change in each developed system.

B. Speech signal acquisition and pre processing

The signals used in this work were acquired by Creative Labs sound board (Sound Blaster model CT4500), with 44.1kHz sampling rate, 16 bits, mono. Each utterance was recorded in .wav file. Silence segments were manually

detected and removed by the *Creative Studio* software (*Creative Labs*, version 3.20.0, 1996). These files have been 4kHz lowpass filtered (4th order Butterworth), leading to the signals submitted to the verification system, that were segmented in 25ms frames using a Hann window

The parameters comparison were performed with a polysyllabic word presenting a wide number of characteristic phonemes of Portuguese, the word "bioinstrumentação" (bioinstrumentation), that was repeated five times for each of the five volunteer speakers. Four of the five utterances were used to derive the reference covariance matrix and the last utterance used as unknown signal.

C. Features

Several computational routines [14] were investigated to the parameter extraction procedure. The LPC parameter (twelve coefficients) was adopted as the primary parameter, from which the other 11 parameters have been calculated through linear/nonlinear conversions using the toolbox Voicebox [15] and specific routines specially developed in Matlab 5.2. The eleven studied parameters were:

- Autocorrelation Coefficients (AC)
- Area Coefficients (AF)
- Area Ratios (AO)
- Complex Cepstral Coefficients (CC)
- Formants frequencies (FF)
- Log area Coefficients (LA)
- Log area ratios (LO)
- Line Spectrum Pairs (LSP)
- Autocorrelation Coefficients of the inverse filter's impulsive response (RA)
- Reflection Coefficients (RF)
- Z-plane autoregressive poles (ZZ)

For each speech-dependent parameter a performance table was obtained, where each row contains the known speaker and each column contains the unknown.

III. Results and Discussion

The result of the twelve studied parameters were organized in tables like the one shown in Table I, where the performance for the LSP parameter can be seen.

TABLE I
PERFORMANCE TABLE FOR LSP PARAMETER

Known Speaker	Unknown speaker				
	1	2	3	4	5
1	**10.2**	192.3	39.4	162.2	126.0
2	54.6	**5.4**	1361.5	61.5	247.7
3	72.2	48.7	**13.7**	57.1	161.6
4	29.9	37.0	60.5	**26.4**	237.5
5	73.3	63.0	130.8	124.4	**2.7**

The system performance can be evaluated by the difference between the values in principal diagonal (corresponding to a correct verification) and the values outside the principal diagonal (corresponding to a wrong verification). If the system is efficient, the element of the principal diagonal presents the minimum value of the correspondent row and column. The systems can be compared one to each other through the ratio between the smaller value outside the principal diagonal (OutD) and the value in the principal diagonal (InD), for the worst case, that is represented by the smaller value in the principal diagonal. The best performance for all the studied parameters in speaker verification system will be the one that exhibits the greater ratio OutD/InD. In this situation, considering that the preprocessing and comparison techniques were the same for all the cases, the best parameter will be got.

Table II presents the result for the parameters that exhibited correct verifications for all the unknown speakers. It was observed that some parameters (formants, magnitudes of z-poles, inverse filter coefficients, area ratios, log area ratios) presented identification errors and for such parameters a performance comparison was not realized.

TABELA II
PERFORMANCE COMPARATION BETWEEN FEATURES

Feature	Performance values/ratio		
	OutD	InD	OutD/InD
RA	27.54	09.87	2.79
RF	36.87	18.38	2.00
AF	36.16	19.39	1.86
LA	45.05	28.97	1.55
CC	36.29	24.80	1.46
LPC	48.64	36.16	1.34
LSP	29.97	26.43	1.13

V. Conclusion

Despite the fact that literature has shown a great number of works using cepstral parameters [5, 8, 11, 16, 17], the comparison realized in the present paper indicates that the parameters correlated with the biometrics characteristics of the speaker are among the best options to the design of speaker verification systems. Thus, the function area parameter seems to be a good choice, although it has just presented as the third best result. Only a more generic study, in the sense of number of speakers and different utterances can really confirm if it is worse than autocorrelation coefficients of the inverse filter's impulsive response or reflection coefficients. It must be mentioned that although all eleven parameters have been derived from the LPC parameters, the linear/nonlinear conversion can lead to a more suitable parameter to the speaker verification system than the primary LPC parameter.

Despite the good performance observed for the verification system shown in Table II, when used with Sambur's technique [13], it couldn't be used in forensic applications [1]. For such cases a probability that unknown speaker is some of the true speakers is desired instead of a distance

value. Then one can conclude that the research must continue in order to develop a more generic system that could be more widely applied.

REFERENCES

[1] C. Champod, D. Meuwly, "The inference of identity in forensic speaker recognition", *Speech Communication*, No.31, pp. 193-203, 2000.

[2] J.M. Naik, "Speaker verification: a tutorial", *IEEE Communications Magazine*, Jan. 1990.

[3] S. Hadjitodorov, B. Boyanov, B. Teston, "Laryngeal pathology detection by means of class-specific neural maps", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 4, No. 1, March 2000.

[4] D.J. France *et al.*, "Acoustical properties of speech as indicators of depression and suicidal risk*", IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 7, July 2000.

[5] S.P. Kishore, B. Yegnanarayana, "Speaker verification: minimizing the channel effects using autoassociative neural network models*",* in *Proc. ICASSP 2000*, Vol. 2, 2000.

[6] K.R. Farrel, R.J. Mammone, "Speaker identification using neural tree networks", in *Proc. ICASSP 94*, Vol. 1, pp. 165-168, 1994.

[7] H. Liou, R.J. Mammone, "A subword neural tree network approach to text-dependent speaker verification*"*, in *Proc. ICASSP 95*, Vol. 1, pp. 357-360, 1995.

[8] D.A. Reynolds, "Speaker identification and verification using gaussian mixture speaker model", in *Proceedings of ICASSP 95*, Vol. 1, pp. 29-32, 1995.

[9] R. Vergin, D. O'SHAUGHNESSY, "On the use of some divergence measures in speaker recognition", in *Proc. ICASSP 1999*, Vol. 2, 1999.

[10] K.R. Farrel, "Text-dependent speaker verification using data fusion*"*, 1995.

[11] B. Imperl, Z. KACIC, B. HORVART, "A study of harmonic features for the speaker recognition"*, Speech Communication*, No. 22, pp. 385-402, 1997.

[12] W.D. Zhang, M.W. Mak, "A two-stage scoring method combining world and cohort models for speaker verification*",* in *Proc. ICASSP 2000*, Vol. 2, 2000.

[13] M. Sambur, "Speaker recognition using orthogonal linear prediction*",* in *IEEE Trans.Acoust. Speech Sig. Processing ASSP*, 1976.

[14] S. Saito, *Speech Science and technology,* 1st ed., Tokyo, Ohmsha, 1992.

[15] M. Brookes, "VOICEBOX: Speech processing toolbox for MATLAB" (current April 18, 2001).
http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[16] S. Furui, "Comparasion of speaker recognition methods using statistical features and dynamic features", *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. ASSP-29, No. 3, June 1981.

[17] K.H. Yuo, H.C. Wang, "Joint estimation of feature transformation parameters and gaussian mixture model for speaker identification", *Speech Communication,* No.28, pp. 227-241, 1999.