TOWARDS AN ACOUSTICAL - ANATOMICAL SYSTEM FOR SPEAKER IDENTIFICATION

Feriel DEBBECHE [1], Nacira GHOUALMI-ZINE [2]
University of Badji Mokhtar, BP 12, Annaba, Algeria
[1] f_debbeche@yahoo.fr
[2] ghoualmi@yahoo.fr

ABSTRACT

A person's identification on the basis of his/her voice presents a real challenge for researchers, due to its intrinsic and extrinsic variability. Several methods have been proposed essentially on the basis of the speaker's modelling. In this paper, we present in a first step a state of the art on the approaches of modelling and a comparative survey. In a second step we propose an acoustical - anatomical system of speaker's identification via a relative modelling. Also, we present the speaker under his/her two acoustical and anatomical facets. To improve the rate of identification and to give a stamp merely biometric to the AIS system, we combine the acoustic features and the anatomical features of the speaker. More precisely the anatomical features of the speaker's ENT sphere.

For this, we combine in a same vector the speaker's acoustical and anatomical features. We also present a model based on the relative approach that consists of presenting a speaker, not in an absolute way, but relatively to a set of recognized speakers. Therefore we define the architecture of the system of identification through introducing our propositions.

Keywords: *Voice, Modelling of the Speaker, Acoustic Features, Anatomical Features, Relative Approach, security.*

1. INTRODUCTION

The speaker's automatic recognition is a domain of the automatic treatment of the speech in which the objective is to determine a speaker's identity by the analysis of his/her voice, in contrast with the speech recognition that consists of the survey of the linguistic content of the state message. Researchs in these domains are closely associated to the computer security.

The speaker's recognition is founded on the variability between speakers and its task is to extract from the speech signal the information in a way to inform on the specificity of an individual: identity, physical features, emotivity, pathological state or regional particularities.

The two pionneer tasks of the Speaker's Automatic Recognition systems (ASR) are the Speaker's Automatic Identification (ASI) and the Speaker's Automatic Verification (ASV) [1, 6, 26].

The Speaker's Automatic Verification is the decisional process permitting to determine, by the means of a vocal message, the veracity of the identity claimed by an individual. In another side, the Speaker's Automatic Identification, topic of this paper, is the process that consists of determining, among a population of known speakers, the person having pronounced a given message.

In the second section of this paper, we describe the speaker under his/her two anatomical and acoustical aspects.

A state of the art on the modelling methods of the speaker is presented in the third section. The fourth section details the relative approach of modelling. In the fifth section, we describe our architecture of the Speaker's Automatic Identification of that bases on the use of a vector combining the acoustic and anatomical features of the speaker.

We finish by a conclusion and perspectives.

2. DESCRIPTION OF SPEAKER

2.1. ANATOMICAL DESCRIPTION

Every person's voice depends on anatomical and behavioural features at the same time.

The vocal device is constituted of structures belonging to the respiratory device and to the digestive device. We split it classically in three steps: [7, 12, 27]

1. The blower: it includes the respiratory musculature, the lungs, and the superjacent pipes.

The blower produces the flux of air that will be the raw material of the vocal production, expired by the lungs and routed by the windpipe toward the larynx.

2. The vibrator: It is about the larynx, that is a tube situated at the superior extremity of the windpipe, to the level of Adam's apple. The column of air produced by the blower is put in vibration under the action of the vocal cords.

3. The resonators: These are mainly the supra-laryngeal cavities, notably the pharynx, the buccal cavity and the nasal pits.

The shape and the volume of these cavities are very variable according to the individuals; it is what explains that every person has a personal and identifiable voice stamp. In addition, the movements of the muscles of the pharynx and the mouth (notably of the tongue) permit fast modifications of the volume and the shape of these resonators, that transform the voice produced by the laryngeal vibration in phonemes, constituent of the articulated speech, by the selective amplification of some laryngeal frequencies.

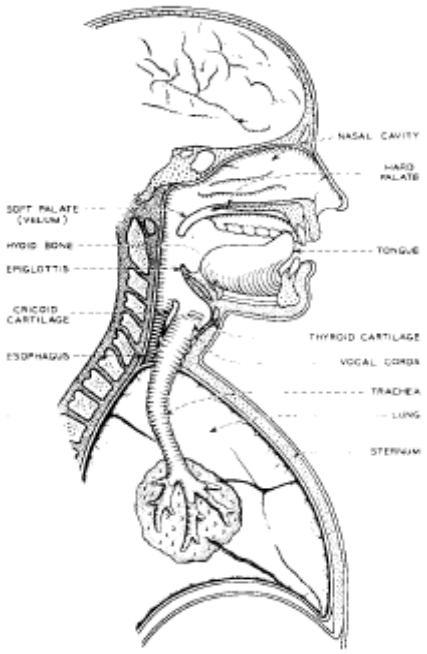

Figure 1: Human Vocal System [7]

The vocal cords are connected horizontally between the thyroid cartilage (Adam's apple at the man) situated at the front and the arytenoid cartilages situated at the back. While making move these cartilages in speaking, we modify the length and position of the vocal cords.

When the person begins to say some words, the arytenoid cartilages stick the vocal cords one against the other, closing the glottis as well.

Under the pressure of expired air, the vocal cords open, then they close again immediately, driving again a rise of the pressure under the glottis.

While opening and closing the glottis during the phonation, the vocal cords free in a jerky way te air stored in the lungs. During a sentence, the speaker modifies hence several times the frequency of vibration of the vocal cords to produce the acoustic vibrations corresponding to different sounds. [7, 12, 27]

2.2. PHYSICAL DESCRIPTION OF THE VOCAL SIGNAL

In addition to the linguistic message serving to the communication between individuals, the signal of speech conveys characteristic information of the person who gave it out as the stamp of his/her voice, his/her way of speaking, his/her emotional or pathological state, etc.

The speaker's characteristic information can be classified in two distinct categories:

- The information of static nature as the spectral parameters characterizing the vocal and nasal pipes, the average and the variations of the fundamental frequency.
- The information of dynamic nature reflecting the phenomena of co-articulation, the formant trajectories as well as the temporal information (speed of elocution, distribution of the pauses).

In this section we speak of acoustic features of the vocal signal that can be defined by 4 main parameters: [11, 23, 31]

1. Intensity: The intensity of a sound corresponds to the amplitude of the acoustic vibration, it characterizes the resonant volume that permits us distinguish between a strong sound of a weak one.
   The vocal intensity especially varies according to the glottis pressure.
2. Stamp: The stamp permits differentiate two sounds of the same height and the same amplitude. It constituted of a set of frequencies called specter. The riches of the specter will permit say that a sound is rich, brillant, deep, etc.
   The stamp is function of the following three criterias: of the conditions of the sticking of the vocal cords, of their thickness and finally of the anatomical features of the resonance cavities (pharynx, mouth and nasal cavities).
3. Pitch: The pitch depends on the frequency of the variation of the acoustic pressure corresponding to the sound.
   It is function of the periodicity of the movement of the lips, that means in practice, of the number of glottis openings per second.
   The pitch depends also on the size of the larynx: the more the vocal cords are long, more the voice is low-pitched.
4. Frequency: It represents the number of air vibrations in one second.

3. THE APPROACHES OF MODELLING (STATE OF THE ART)

A system of Automatic identification of the Speaker is summed up in the sequence of three main processes: the speech signal parameterization, the modelling and the decision.

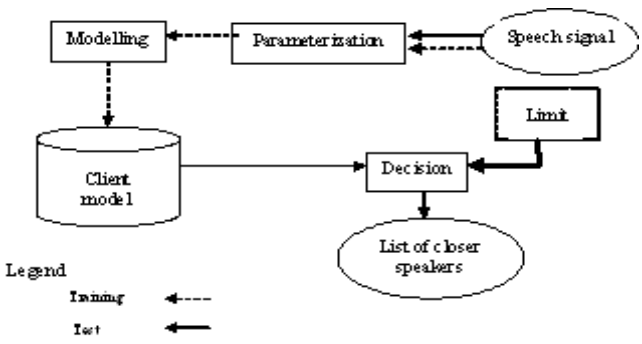

Fig ure 2: ASI System Structure.

The researchers in ASI concentrate essentially on the second process.

In this axis, numerous approaches have been proposed: vectorial approach, connexionnist, predictive, statistical, relative, etc. Of this large panel, only the statistical approach remains at the first place of the systems of the speakers' automatic recognition of the recent years.

However, the performances in these approaches deteriorate considerably if the data of training are insufficient. However, in most applications, the phase of enrollment must be very short (of the order of some seconds of speech).

To remedy this problem, new approaches appeared of which the relative approach, that modelises the speaker, not in an absolute way, but relatively to a well trained set of speakers. In an other word, every speaker is represented by his localization in an optimal space of neighbors. [14]

A state of the art and a comparative survey on the existing methods of modelling of the speakers are summarized in the Table 1.

4. RELATIVE APPROACH

The principle of identification of the speakers by the relative approach is based the representation of every speaker by its location in an optimal space of neighbors.[13, 15, 17, 20, 9]

This approach came to compensate one of major disadvantages of statistical approach (remaining at the first place systems of the speakers' automatic recognition of the recent years.) which is the big number of the training data required for obtaining good performances.

From this perspective, the systems are generally divided in three modules: in the first, a space of representation is constructed while the second is dedicated to the new speaker location in this space. In the last module, a test of recognition is done. [17, 20]

A speaker's representation by its localization in the space of the speakers presumes therefore more of the speakers will be different; more the Euclidean distance in this space grows.

4.1. CONSTRUCTION OF THE REPRESENTATION SPACE

The construction of a space of representation or eingenspace can be done either by data analysis methods (PCA, PPCA, LDA), Maximum Likelihood EingenSpace (MLES), Hierarchical Clustering or by selection (Genetic Algorithms).

4.2. LOCALIZATION OF THE SPEAKERS

The localization of the speakers as for it is done by orthogonal projection on the eigenspace constructed previously, by MLED or by the anchor models.

4.3. DECISION

The intuitive representation of a speaker by its localization in a space of representation presumes that more speakers are similar more their points of projection are near and the distance between them is small. Therefore, to exploit the notion of neighborhood and to value the proximity in the space, a metrics is used between the coordinates of the speakers of training and the test.

5. PROPOSED ARCHITECTURE

The information contained in the signal of the speech can be divided in several levels: acoustic level, prosodic, phonetic, idiolectal, dialogal and semantic. The level that was the more exploited, by its simplicity of use, in the systems of ASI or broadly speaking in the systems of ASR is the acoustic level.

To reinforce the identification rate and to give a purely biometric stamp to the system of ASI, an interesting perspective consists of using, next to the acoustic features, the speaker's anatomical features.

The anatomical parameters of the vocal device are very different according to the individuals this what justifies their use in a system of identification. The table.2 and figure.3. below [12] illustrate this difference in representing the typical values of the glottal geometry.

Table 1: Compartive Chart.

Approaches	Advantages	Disadvantages
DTW **[4, 8, 30]**	• Very fast • showing relatively good performances	• Used exclusively in text dependent mode • Very sensitive to the quality of alignment and to the choice of the starting point.
QV **[18, 19, 28, 30]**	• Can be applied in dependent to text mode or in independent to text mode	• The speed and the performances depend strongly on the size of the codebook: the more the size of the codebook increases, better are the performances; nevertheless, the process becomes much slower
Connexionist Approaches **[21, 22]**	• Good performances	• Complexity of training. The addition of a new client requires the retraining of all models.
Predictive Approache **[2, 9, 10, 16]**	• Taking into account the dynamic information transported by the speech signal.	• The performance is still not enough for practical use.
Relative Approache **[13, 15, 17, 20, 29]**	• The modelling of a new speaker is not in a absolute manner but relatively to a set of well trained speakers.	• The identification rate depends on the quantity of the training data (best rate if we have less data)
HMM **[5, 26]**	• Taking into consideration the temporal aspect of the speech signal. • Excellent results in text dependent mode.	
GMM **[24, 25]**	• Good performances in text independent mode.	• The major disadvantage is the quantity of training signals required for a good evaluation of the models parameters.
SMSO **[3, 16]**	• implementation simplicity • Efficient on short duration.	• The local variations are not taken into account by the models.
SVM **[29]**	• Good results.	• Long time of training.

Table 2: Typical Values of the Glottal Geometry.

Length of vocal cords	L_g	14 – 18 *mm*
Thickness of vocal cords	d	5 – 10 *mm*
Glottal distanc	h	0 – 3 *mm*

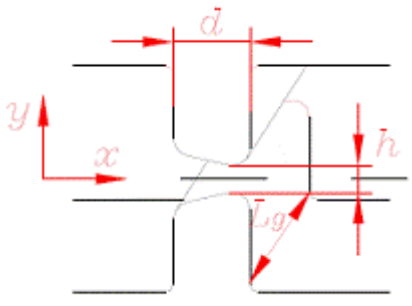

Figure 3: Geometry of the Glottis

Therefore, the architecture that we propose combines the acoustic and anatomical features of the speaker in order to construct an acoustical - anatomical system of ASI.

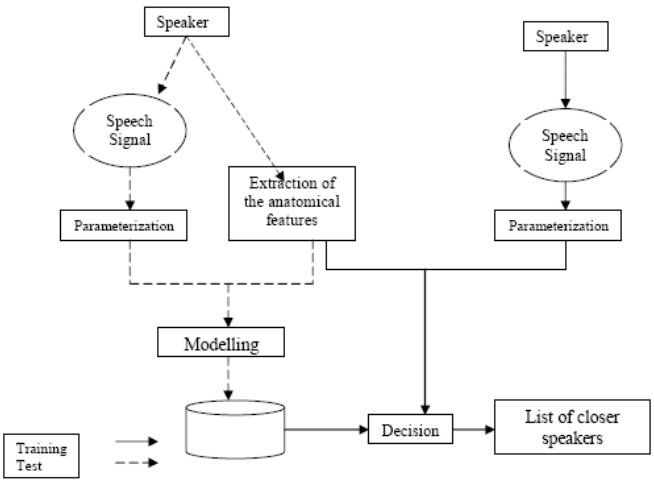

Figure 4: Structure of the Acoustical-Anatomical System.

6. CONCLUSION

The speaker's automatic identification is the result of a synergy of a tripod englobing a process of parameterization of the vocal signal given out by the speaker, of a process of modelling and of a phase of decision.

In this paper, we present a state of the art and a comparative survey on the different approaches of modelling of the speaker.

To give a stamp purely biometric to our system of identification, we propose architecture based on the use of a vector combining the acoustic and anatomical features of the speaker.

Therefore, our acoustical - anatomical system uses this vector as entry to our model based on the relative approach.

As perspectives, we have to detail our approach of modelling namely: to choosing a method of construction of the eigenspace, a method of localization of the speakers and a metrics for the decision.

The construction of a corpus and the implementation of our architecture are in the course of realization.

REFERENCES

[1] Atal B. S., "Automatic recognition of speakers from their voices," IEEE transactions, volume. 64(4), pages. 460-475, 1976.

[2] Bimbot F. Mathan L. De Lima A. Chollet G., "Standard and target driven AR-Vector Models for speech analysis and speaker recognition," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages. 5-8, San Francisco (USA), 1992.

[3] Bimbot F. Magrin Chagnolleau I. Mathan L., "Second-order statistical measures for text-independent speaker identification," Speech Communication, volume. 17(1-2), pages. 177-192, Aôut 1995.

[4] Booth I. Barlow M. Watson B., "Enhancements to DTW and VQ decision algorithms for speaker recognition," Speech Communication, volume. 13 (3-4), pages. 427-433, Décembre 1993.

[5] De Veth J. Bourlard H., "Comparison of hidden Markov model techniques for automatic speaker verification," Workshop on Automatic Speaker Recognition, Identification, Verification, pages. 11-14, Martigny (Suisse), Avril 1994.

[6] Doddington G. R., "Speaker recognition Identifying people by their voices," IEEE transactions, volume. 73(11), pages. 1651-1664, 1985.

[7] Flanagan J., Speech Analysis Synthesis and Perception, 2nd ed, New York and Berlin: Springer-Verlag, 1972.

[8] Furui S., "Cepstral analysis technique for automatic speaker verification," IEEE Transactions Acoustics, Speech, and Signal Processing (ASSP), volume. 29(2), pages. 254-272, Avril 1981.

[9] Grenier Y., "Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur," IXèmes Journées d'Etudes sur la Parole (JEP), pages. 163-171, Strasbourg (France), 1980.

[10] Griffin C. Matsui T. Furui S., "Distance measures for text-independent speaker recognition based on MAR model," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages. 309-312, Adélaide (Australie), 1994.

[11] Homayounpour M M. Chollet G., "Performance comparison of some relevant spectral representations for speaker verification," Workshop on Automatic Speaker Recognition, Identification, Verification, pages. 27-30, Martigny (Suisse), Avril 1994.

[12] Kob M., "Physiologie des lèvres et des cordes vocales," Hôpital de phoniatrie, othophonie et dysfonctionnements de communication, Université d'Aix la Chapelle – RWTH, Allemagne.

[13] Kuhn R. Nguyen P. Junqua J-C. Goldwasser L. Niedzielski N. Fincke S. Field K. and Contolini M., "Eigenvoices for speaker adaptation," ICSLP, 1998.

[14] Kuhn R. Nguyen P. Junqua J-C. Goldwasser L. Niedzielski N. Fincke S. and Field K., "Eigenfaces and eigenvoices: di-mensionality reduction for specialized pattern recognition," MMSP, 1998.

[15] Kuhn R. Nguyen P. Junqua J-C. Boman R. Niedzielski N. Fincke S. Field K. and Contolini M., "Fast Speaker Adaptation in Eigenvoice Space," ICASSP, 1999.

[16] Magrin Chagnolleau I. Wilke J. Bimbot F., "Further investigation on AR-vector models for text-independent speaker identification," International Conference on Acoustics, Speech,

and Signal Processing (ICASSP), pages. 401-404, Atlanta (USA), 1996.

[17] Mami Y. Charlet D., "Identification des locuteurs par regroupement hiérarchique ascendant et modèles d'ancrage," XXIVèmes Journées d'Étude sur la Parole, Nancy, 24-27 juin 2002.

[18] Mason JS. Oglesby J. Xu L., "Codebooks to optimise speaker recognition," European Conference on Speech Communication and Technology (Eurospeech), pages. 267-270, Paris (France), 1989.

[19] Matsui T. Furui S., "Comparison of text-independent speaker recognition methods using VQ-distorsion and discrete-continuous HMMs," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages. 157-160, SanFrancisco (USA), 1992.

[20] Nguyen P[1,2]. Kuhn R[1]. Junqua J-C[1]. Niedzielski N[1]. Wellekens C[2]., "Voix Propres: Une représentation compacte de locuteurs dans l'espace des modèles," [1]Speech Technology Laboratory, Santa Barbara, Californie. [2]Institut Eur_ecom, Sophia-Antipolis, France. CORESA'99 14-15 Juin 1999.

[21] Oglesby J. Mason JS., "Optimisation of neural models for speaker identification," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages. 261-264, 1990.

[22] Oglesby J. Mason JS., "Radial basis function networks for speaker recognition," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages. 393-396, Toronto (Canada), 1991.

[23] Reynolds D A., "Experimental evaluation of features for robust speaker identification," IEEE transactions Speech Audio Processing, volume. 2, pages. 639-643, 1994.

[24] Reynolds D A., "Speaker identification and verification using gaussian mixture speaker models," Speech Communication, volume. 17(1-2), pages. 91-108, 1995.

[25] Reynolds D A. Quatieri T F. Dunn R B., "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing (DSP), a review journal – Special issue on NIST 1999 speaker recognition workshop, 10(1-3), 2000.

[26] Rosenberg A E. Soong F K., "Recent research in automatic speaker recognition," Advances in speech signal processing, 1991.

[27] Roublot P. "Analyse comparative subjective et objective de la voix avant et apres bloc interscalenique du plexus brachial," Universite Henri Poincare, Nancy I, 2003.

[28] Soong F K. Rosenberg A E. Rabiner L R. Juang B H., "A vector quantization approach to speaker recognition," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages. 387-390, Tampa (USA), 1992.

[29] Vapnik V., The nature of statistical learning theory, Spring-Verlag, New York, 1995.

[30] Yu K. Mason J S. Oglesby J., "Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation," IEE vision, image and signal processing, Berlin (Allemagne), 1995.

[31] Zwicker E. Feldtkeller R., "Psychoacoustique," CENT/ENST, collection technique et scientifique des télécommunications,Mason Paris, 1981.