# AN OBJECTIVE AND A SUBJECTIVE APPROACH OF SPEAKER RECOGNITION

**Elisabeth Lhote, Laura Abou Haidar**

**Laboratoire de Phonétique - 30 Rue Mégevand - 25000 Besançon - France**

## ABSTRACT

We consider speaker recognition as an integration level in the transfer process from production to understanding. In tackling speaker's recognition from the point of view of proximity between several speakers, we chose two complementary approaches : a "descending" approach, that allows extracting objective elements in both auditory and acoustic analysis , in order to associate voices unknown from the experimentor ; a "rising" approach, that allows bringing to light objective criteria for the characterization of vocal proximity between speakers close at a genetic, acoustic or auditory level.

## 1. INTRODUCTION

Speaker recognition is considered here as a key process in speech recognition. The listener who recognizes someone by his voice resorts to various treatment mechanisms : for a global treatment, he refers to discourse analysis ; for a local treatment, he selects acoustic characteristics which, memorized, become attributes characterizing one speaker. From the point of view of the listener, the two treatment models are associated and it is difficult to know whether one of them influences the other and how does the listener proceeds in distinguishing the two. It is often said that this approach is subjective. In fact, the recognition by the listener is done in real time : as soon as he hears the first words on the phone, he usually knows who is calling him amongst people he knows. This observation brings to the front, in daily practice, an ability to select and associate vocal attributes with a known person. However, sometimes, doubt disturbs recognition. The listener hesitates between two people. We are interested by this situation in as much as the listener's recognition system is not sufficient. We decided to tackle speaker's recognition from the point of view of proximity between several speakers.

## 2. HYPOTHESIS

Our hypothesis is the following : whatever the discourse of the speaker may be, and whatever his emotional state, the neuro-articulatory and neuro-phonatory mechanisms which command and control the speech neurolinguistic programming are constant. This does not mean that the way we produce a syllable remains the same for each speaker, but that a neurolinguistic invariability remains as long as a pathological affection does not alter the voice.

## 3. EXPERIMENTATION

The experimentation focuses on comparison between different speakers, according to two complementary approaches : one called "descending", the other "rising".

In the first approach, we tried to associate unknown voices that had been recorded, with models. This "descending" approach allowed us on the one hand to extract objective elements in both auditory and acoustic analysis ; on the other hand, we were better able to estimate the notion of proximity between voices.

In the "rising" approach, we selected speakers close in age, with family ties, with similar ways of talking, and having voices which are similarly confused on the phone. Then we tried to bring to light objective criteria allowing to characterize vocal proximity.

### 3.1. Descending approach

The first group was constituted of five speakers : S.A, S.B, S.C, S.D, S.E, and the second of twelve, among whom could be found the five speakers of the first group. In this case, we had to match voices of speakers reading a text varying between 2 to 5 minutes, of which only some sentences were produced by speakers belonging to both groups. The auditory analysis consisted of a systematic analysis of discourses at a phonetic level.

#### 3.1.1. Global parameters

The global parameters which were the most pertinent were rhythm and intonation. In order to better bring them to light, we performed a simultaneous auditory analysis of two voices producing for instance the same sentences. A correlation between auditory and acoustic analysis allowed us to bring to the fore front ways of talking that are close and distant (FIGURES 1 & 2).

#### 3.1.2. Local parameters

Afterwards, local analysis parameters were extracted by spectral analysis : a systematic analysis of formant trajectories in key sequences allowed us to put together or to separate some speakers (FIGURES 3 & 4).

The final results obtained with the help of this double analysis : local and global, auditory and acoustic, are positive and show the efficiency of this approach in discovering unknown links between voices and speakers.

### 3.2. Rising approach

In this case, speakers are known by the experimentor. The corpus is elaborated in order to bring to the light formant structures visible in key words or key syllables.

#### 3.2.1. Acoustic proximity

Thirteen speakers produced the following text twice :

*"Tu sais, pendant les vacances à la montagne avec Jean, il y avait de ces tourbilllons! Les tourbillons étaient trop forts!"*

The selected syllable was [jõ] in "tourbillons". The results of this analysis [4] showed a greater or lesser variability of slopes depending on the speaker. And particularly they allowed us to select 2 speakers whose slopes were very close. We recorded these two speakers again, and we asked them to vary their voice. One sentence :

*"Les tourbillons de Lyon"*

was produced 40 times by each of them : 10 times in a normal voice, 10 times whispering, 10 times shouting, 10 questioning. We tried to extract a cue characterizing either the articulatory movement or an articulatory invariability.

- [bijõ]

The slope analysis of the two syllables [bijo] in different voices did not permit differentiation between the two speakers.

- [ɔ]

We noticed that the following cue :

[F4 - F3]

could be dependent of speaker's vocal behaviour : when converting these frequential values in tones, we noticed that this tonal cue seems to be an element that could characterize speakers' vocal behaviour :

* in the first speaker, the value of this tonal cue was : 3 tones, whichever voice was used ;
* in the second speaker, a variation of this cue was situated between two and three tones depending on the type of voice.

It is important to underline that from an auditory point of view, these two speakers don't have the same voice, even if the acoustic analysis shows a very close proximity.

#### 3.2.2. Genetic proximity

We analysed three sisters' voices Y, L, N, two of which are often mixed up on the phone (L & N). We tried to find whether acoustic cues linked to formant

transitions gave an explanation of this proximity.
The tested sentence was the following :

*"Il y avait de ces tourbillons! Les tourbillons étaient trop forts!".*

The key syllable was [jð] in "tourbillons". We selected the slope of F2 between [j] and [ð] and calculated it into tones. We think that this cue should contribute to define the velocity of the articulatory movement. We obtained the following results (FIGURE 5) :
Number of tones for a 40 ms interval :

L --> 6 tones
N --> 4 1/2 tones
Y --> 4 1/2 tones

Other experiments showed us that this tonal slope cue of the first three formants can be steady in some speakers production and unsteady in others when they change from normal voice to shouting, whispering, questioning. We were expecting to find the same slope values in L & N, who are often mixed up on the phone ; in fact, we didn't. We deduce from this result that **results obtained at the auditory level can be different from those obtained at the acoustic level.**

## 4. CONCLUSION

After having tested the relation existing between the auditory appreciation of a voice and its acoustic analysis - global and local -, we extracted the following points :

- Two voices auditorily close can be distant acoustically and vice versa ; that is why it is important to associate the two approaches which should be considered complementary.
- If we are looking to characterize the articulatory movement velocity, it is useful to take into account the formant 4 and to use slope tonal variations.
- However, it should be noted that what appears to be necessary - during the rising approach - to the differenciation between two speakers is not necessarily sufficient to succeed in identifying a speaker from others during a descending approach.

**In speakers recognition, as well as in speech recognition, a systematic correlation between the different analysis levels is necessary, in order to avoid favoring cues which belong to a unique analysis level.**

## 5. REFERENCES


[1] BOOMER D.S., LAVER J.D.M. (1968) : Slips of the tongue. *British Journal of Disorders of Communication*, 3, 1-11.

[2] HARDCSTLE W.J. (1976) : *Physiology of Speech Production*, Academic Press, London.

[3] LENNEBERG E.H. (1967) : *Biological Foundations of Language*. New York : John Wiley and sons.

[4] LHOTE E., ABOU HAIDAR L. (1990) : Speaker verification by a vocal proxemy cue. *ESCA Tutorial and Research Workshop on Speaker Characterisation in Speech Technology*, Edinburgh, 149-154.

[5] LIENARD J.-S. (1989) : Variabilité, contrainte et spécificité de la parole : un cadre théorique. Actes du Séminaire : *Variabilité et Spécificité du locuteur*, CIRM - Marseille, Luminy.

[6] MAC NEILAGE P.F., DE CLERK J.L. (1969) : On the motor control of coarticulation in CVC mono-syllables. *Journal of the Acoustical Society of America*, 45, 1217-1233.

[7] OHMAN S.E.G. (1966) : Co-articulation in VCV utterances : spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151-168.

[8] PASCAL D. (1989) : Etude de la similarité des voix masculines : corrélation entre mesures physiques et structure perceptive. Actes du Séminaire : *Variabilité et Spécificité du locuteur*, CIRM - Marseille, Luminy.


## 6. FIGURES

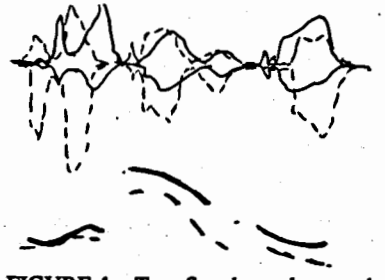

FIGURE 1 - Two Speakers close at the rhythmic and melodic level
*Sentence : "C'est d'accord ou quoi ?"*



FIGURE 2 - Two Speakers distant at the rhythmic and melodic level
*Sentence : "C'est d'accord ou quoi ?"*

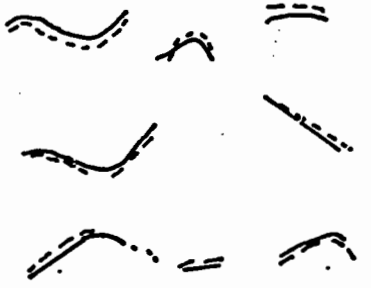

FIGURE 3 - Formant trajectories of two close speakers

FIGURE 4 - Formant trajectories of two distant speakers

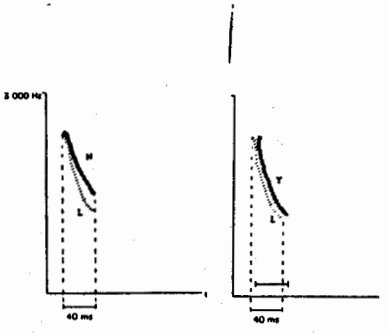


FIGURE 5 - F2 transition slope in the syllable [ijð]