



INTER AND INTRA-SPEAKER VARIABILITY OF FRENCH PHONEMES ADVANTAGES OF AN EXPLICIT KNOWLEDGE BASED APPROACH

Jean-François Bonastre, Henri Méloni

LIUAPV, Avignon, France

Abstract— This article deals with a knowledge based approach in connexion with automatic speaker recognition and, in particular the study of specific spectral parameters of the speaker.

Key words — Identification, Spectral, Voice, Knowledge, Analytic.

1. INTRODUCTION

We have focused our work on the speaker's specific information contained in short term spectra. In this context, we have based our work on discriminant capacities of each phoneme, combined with spectrum comparison tools. Then, we have quantified variations of phonemes spectrum representation and tried to reduce their influences. In this respect, we have used an analytic based approach which enables the use of certified knowledge.

2. SPEECH VARIABILITY

Speech variability isn't homogeneous. It is made of four types of variability : random variability, context variability, inter-speaker variability, and language variability [6]. In a data processing system, we have introduced an additional variability cause : processing variability . Of course, all causes of variability do not disturb the ASR system, because all the ASR system is based on inter-individual variability. We have been seeking easy to use and independent indexes (as far as possible from the duration, grammatical composition, etc.). Random variability, whose shape is unpredictable but results quantifiable, reduces the dynamic of an inter-speaker similarity calculation without casting doubt over the index we have used. Data processing variability leads us to the same conclusion. On the contrary, context variability (in particular coarticulation effects) causes major problems. Spectral appearance, the place of formants, the duration and the stress of a phoneme are influenced by the context [5] These variations remain one of the great problems of systems allowing utterances of variable nature. On the contrary, the variability of the language is not significant in practical applications.

3. COMPARISON PARAMETERS

Spectral shapes comparison is quite difficult to achieve. There are still tricky details, like context variability, energy levels, the influence of parameterisation and representation of spectral vectors method. There is not a single answer. We want to create various comparison parameters, each one giving evidence of a specific piece of information. At our disposal we have a whole of comparison parameters that enables us to scan all spectrum characteristics. [3]. Each determined spectral distance can as well fit the type of spectral representation (24 channels Mel or 128 linear channels), the type of phoneme (scanning formants for vowels, remote peaks of energy for the /l/, energetic adjustment, etc.) as the dynamic we aim at (selection of the n worst channels, use of an increasing threshold, etc.). The optimum comparison between both

segments of the signal is made through a selection of various distances, well fitted to spectral representation.

4. ASSESSMENT OF VARIOUS PARAMETERS

The wide range of choice of discriminant indexes produces a new difficulty when systems start to decide : how matching a $x\%$ discrepancy in the size of a formant with the duration discrepancy of a phoneme or with the FO evolution ?

In the process of time the comparison of indexes is made thanks to *F-ratio* [4] :

$$F = \frac{\text{Variance of the means of different speakers}}{\text{Average of intra - speaker variance}}$$

The *F-ratios* provide an assessment of relative performances of indexes (a posteriori and based on a great number of analysed elements). But during a short term decision - based on many reliable indexes and on few analysis elements - the *F-ratio* does not make possible to measure the significance of an index in relation to an other one, and to determine whether the information is sufficient to generate a decision. Moreover, such a criterion requires a great number of samples for each individual in order to assure fair results (the means are calculated speaker by speaker, which requires for every one of them several tens of samples).

Other evaluation criteria may bring a solution, by considering the intra and inter-speaker variation space of each index [1 ; 7 ; 8]. Such an approach enables taking multi-indexes decisions with a parametrable safety margin. However it requires a previous study of each index for modelisation of spaces variation. The criterion of evaluation we suggested gave great importance to the ratio inter-speakers variability by intra-speaker variability without taking into account the discrepancy between variation spaces.

In order to solve this question, we have tried a new technique using the intra and inter-speaker whole means and also standard deviation. This method may emphasize the influence of the means in comparison with scattering factors [3]. We suggest the pertinency criterion σ_m combining the ratio intra and inter-speaker standard deviations with the difference of inter and intra-speaker means.

Tough we have got the formula :

$$\sigma_m = \sqrt{\frac{\sigma_{inter}}{\sigma_{intra}}} \cdot (m_{inter} - m_{intra})$$

(with σ_{intra} standard deviation and m_{intra} global average intra-speaker, σ_{inter} standard deviation et m_{inter} global average inter-speaker).

To visualise the practical capacities of a system based on these techniques, we have also calculated the number of phonemes necessary to the speaker's identification, according to the safety margin selected. This corresponds to the number of cases that ensures a split between intra and inter-speaker variation spaces.

The corresponding formula is :

$$n \geq (a^2 \cdot \left(\frac{\sigma_{intra} + \sigma_{inter}}{M_{inter} - M_{intra}} \right)^2)$$

(a is function of the selected safety margin ; a = 2 for a 95% coefficient ; a = 3 for 99,7%).

Notes :

- Issued values define the most favourable case : the use of many cases belonging to a same phoneme (redundancy of information), for each case we consider that the impersonator is the closest speaker to the individual, etc.
- During the recognition tests, identification is supposed to be successful if - concerning a specific speaker - the value in relation to the index is strictly inferior to other speaker ones' of the base.

5. SPECTRAL INFORMATION : VARIABILITY AND SPECIFICITY OF THE SPEAKER

We suggest to classify the individual characteristics contained in short term spectra of French and to quantify variations allotted to spectral variations of various phonemes.

1. DESCRIBING THE EXPERIENCE

We have extracted a corpus from the BDSON. A corpus made of an utterance labelled as "la bise et le soleil", lasting about 50 seconds and pronounced by 22 speakers. This corpus is made of a single text.

In order to match different parameterization methods of the signal, we calculate spectra showing phonemes of three different ways : through a FFT by a 14 coefficients and a 21 coefficients linear prediction. Spectrums have been modeled of two different ways : 24 values distributed along a scale of Mel and 128 values distributed along a linear scale.

In a precise context, for each spectrum we measure the distance by the corresponding reference of a speaker (intra-speaker distance). On the same way, we calculate the distances by all the others speakers (inter-speakers distances). This works is carried out for each spectrum that characterises a phonetic segment of the utterance. Then we estimate the average distance of a spectrum by the corresponding speaker's reference (so, the standard deviation). The same kind of calculation is undertaken in order to estimate the inter-speaker average mean (mean of distances between each spectrum of the signal and other speaker's references) and inter-speaker standard deviation. We get such results for each studied comparison parameters and for each parameterization method of the signal.

As it was difficult to distinguish learning corpus and recognition corpus, the first spectrum we met in a specific context is stocked as reference. The following ones are registered in the recognition data base.

2. INFLUENCE OF THE SIGNAL PARAMETERIZATION METHOD

There are many ways to change a rough speech signal into a spectrogram. One technique consists in contemplating the very speech signal as any signal and in processing it to a regular way, for instance through a FFT. Other solutions are based on human anatomy and in particular on organs and strategy that make the speech possible. What we wanted to know was if a parameterisation was more accurate and if it give best results or agreements (good visualisation of information capacity for the expert) to justify a possible additional calculation resources. We have assessed the accuracy of the comparison index COMP with the σ_m criterion. These results are assessed for a FFT analysis and a 14 and 21 coefficients modelisation. References are obtained according to the context (one reference for each phonemic triplet and each speaker).

Phoneme	Nb tests	Parameterization		
		FFT	LPC 14	LPC 21
Vowels	1026	148	214	255
Fricatives	259	441	605	595
Occlusives	390	165	207	211
Liquids	356	133	170	183
Semi-V.	171	179	258	262
Total	2202	185	257	264

Table 1 : Table showing the number of intra and inter-speaker measures for each phonetic category and the discriminant power of the COMP index estimated with the σ_m criterion according to the parameterisation method we have used.

Results described in the table 1 show the influence of a production model as the LPC model on comparison indexes since we get an average gain superior to 30% (using the COMP index). On the contrary, best results obtained through the change of a 14 coefficients model for a 21 coefficients model do not justify the increase in calculations.

Notes :

- We underline that comparison parameters are finalized thanks to expert observations, also based on LPC spectrograms.
- It is difficult to split up spectral representation methods from comparison index parameters are specially determined for a representation, or at the least, optimized for a kind of spectral vectors.

3. INFLUENCE OF THE CONTEXT

The main problem ASR systems have had to face, using spectral comparison as selective criterion, consists in coarticulation phenomenon. These ones induce a great spectral variability of the phonemes. This very variability generated by the coarticulation is directly linked with the context of the phonemes. Within the ASR system context we have to measure and to reduce as far as possible the influence of these phenomenon.

Therefore we have tried to know to what extent the immediate past and future of the context alters the spectral shape of a phoneme (it is a long term influence alteration, but because of complexity reasons we won't work on phonetic n-uplets superior to triplets).

We have calculated the phonemic parameters selectivity by the $\mathcal{O}m$ criterion, by the percentage of right identification for one case, by the number of cases necessary to split off intra and inter-speakers variation spaces with a 95% safety margin. We have tried to get these results through different methods taking into account the immediate phonemic context :

- use of the past and next context : one reference is registered for each phonemic triplet.
- only one reference for each phoneme.
- use of the past context : one reference related to a phonetic doublet which is made of the previous phoneme and the phoneme we have studied.
- use of the next context : one reference by doublet.

On table 2 we may notice the influence of the context on the outcome, as well in terms of right identifications (from 49% to 20%) as in number of cases necessary to get a 95% safety margin (from 10 to 60). Phonemes discriminant capacities -measured with the $\mathcal{O}m$ index - have about 50% variation according to the way the context is taken into account.

Few results :

- as it could be expected, the use of a reference by a triplet produces really better results than any other proposition.
- Comparison of results produced by past and future contexts is favourable to future context. As it could be expected, the anticipation of a sound has more influence on the phoneme's acoustic appearance than articulatory constraints previous phoneme.
- There is a major influence of the context on liquids and semi-vowels when only one reference by phoneme is used. Parameters based on the study of vowels are slightly influenced by the context variability (a 65% fall of $\mathcal{O}m$ criterion). Fricatives appear to be less influenced by the coarticulation. Occlusives remain on the average level, sensible to the context variation but getting reasonable results.

6. COMPARISON WITH A "BLIND" METHOD

The technological choices we have made give advantage to a knowledge based approach and enable us the use of the explicit knowledge of the phoneme and its context within spectral comparison parameters. Such an approach, despite the advantages we have demonstrated in the previous paragraph, requires a too difficult localisation process of the phonemes. We want to match our choices with regular technique, close to vector quantification, in order to prove that analytic knowledge based approach advantages justify difficulties of localisation.

In that way, we have kept the previous corpus, extracted from the BDFON. By the same way as previously we have also carried out the acquisition of the references : one reference by phonetic triplet and by speaker and obtained from the standard labelling given with the corpus. We have kept the signal's parameterization method (LPC-21) and the general methodology of extraction of statistics information. On the other hand, the label of a signal unit with stocked references is carried out according to a blind method. Concerning a speaker, the mark given to the unit is estimated by comparison between

the signal unit and each speaker's references, then the result produced means the minimum distance for the very speaker.

Phoneme	nb. tests	$\mathcal{O}m$	id%	nb. 95%
One reference for one phonetic triplet				
Fricatives	259	596	57%	7
Occlusives	390	211	47%	12
Liquids	356	184	38%	14
Semi-vow.	171	263	57%	6
Vowels	1026	226	50%	8
Total	2202	265	49%	10
One reference for one phoneme				
Fricatives	1081	410	31%	14
Occlusives	1265	100	21%	56
Liquids	1304	40	14%	430
Semi-vow.	286	102	30%	100
Vowels	3579	81	18%	62
Total	7515	122	20%	60
Use of the next context				
Fricatives	538	416	38%	14
Occlusives	672	118	30%	38
Liquides	794	114	25%	52
Semi-voy.	218	184	50%	20
Vowels	2054	132	35%	25
Total	4276	168	34%	30
Use of the past context				
Fricatives	577	542	47%	9
Occlusives	611	182	39%	16
Liquids	917	136	28%	24
Semi-vow	221	240	53%	7
Vowels	1980	160	37%	16
Total	4306	220	38%	17

Table 2 : for each phonetic reference, with the *COMP* criterion and according to the way we consider the context : number of tests achieved, $\mathcal{O}m$, percentage of right identifications with an occurrence, number of occurrences necessary to get a 95% safety margin.

Few remarks sum up the conditions of the experiment :

- signal parameterization and spectral representation methods remain unchanged (LPC-21 associated to a 24 channels Mel scale and a 128 channels linear scale).
- the learning phase is carried out with an explicit manual labelling (one reference per speaker and per different phonetic triplet).
- One test unit is defined by the manual labelling (selection of framework showing the unit through the label) but without keeping any information about its characteristics (nor knowledge of the phoneme neither of the context).
- The mark given to the test unit, for a specific speaker, corresponds to the minima obtained distance by comparing the test unit to each reference belonging to the speaker.
- The spectral comparison parameter *COMP* is optimized according to the phoneme and the context. In the context of this experiment, we take into account the knowledge of the reference and of the context in which it was obtained, without regards to the nature of the unit.

pho	explicit informations				"blind" comparisons			
	id%	σ s/i	Mi-Ms	p:95	id%	σ s/i	Mi-Ms	p:95
f	42%	46%	397	7	26%	98%	63	16
s	59%	30%	576	4	29%	111%	70	18
v	50%	63%	261	4	31%	85%	64	19
z	64%	106%	209	7	39%	111%	82	11
p	29%	99%	143	18	18%	93%	38	66
t	41%	82%	215	11	20%	99%	45	42
k	33%	119%	136	26	21%	104%	48	40
b	58%	64%	255	5	23%	95%	64	31
d	53%	68%	266	6	30%	103%	63	26
m	62%	81%	199	5	44%	110%	94	13
l	34%	92%	171	15	26%	103%	55	30
r	46%	79%	176	13	25%	106%	54	36
j	84%	71%	302	2	41%	103%	85	10
a	54%	71%	221	6	28%	104%	63	27
ɔ	50%	102%	161	8	24%	99%	60	29
ε	60%	66%	218	3	23%	100%	56	27
o	70%	95%	159	11	16%	101%	50	52
e	51%	68%	291	6	24%	96%	57	30
ə	29%	63%	146	14	19%	88%	51	35
i	48%	97%	192	6	21%	100%	50	32
y	43%	41%	258	6	15%	98%	36	47
u	48%	59%	228	9	20%	103%	44	54
ã	62%	52%	182	5	21%	88%	56	33
õ	61%	86%	177	4	22%	101%	52	33
Tot	49%	67%	216	10	26%	101%	60	27

Table 3 : Comparisons of identification outcomes with the knowledge of the phoneme and its context and using the "blind" method (with explicit learning) ; percentage of right identifications with one occurrence , inter/intra-speaker standard deviation ratio, difference between inter and intra-speaker means and number of phonemes necessary to a identification with a 95% safety margin (all based on COMP criterion).

The comparison of outcomes obtained by the method of phoneme explicit knowledge and its context, and by a blind method is clearly favourable to the analytic knowledge based approach (Table 3). With only one occurrence we get a 200% in right identification percentage and 300% in number of necessities occurrences for a 95% safety margin.

We observe the same change for all phonemes : inter and intra-speaker variations of same amplitude and standardization of inter-speakers and intra-speaker scores (very sharp reduction -about 300%- of the difference between inter and intra-speaker means).

7. CONCLUSION

We have first given prominence to the information specific to the speaker transmitted by short term spectra embodying sounds of French language. The influence of the surrounding context on phoneme's spectral appearance has seemed to be the source of variations hiding speaker's specificity. Once effects have been measured, we intended to define techniques that minimize the importance of such a phenomenon by taking clearly into account the observed phoneme and its context.

This work enable us to demonstrate that a small number of phonemes (less than 10) was enough to reach successfully, amidst a whole of persons, the most resembling speaker to an individual characterized by a vocal utterance. However, the methodology requires a free will participation of the speakers and a similar environment during the learning phase and recognition tests.

The comparison between the explicit approach (using part of available symbolic information) and blind techniques has demonstrated that analytic knowledge based methods overmatched the other one. Such a result makes profitable the cost of analysis elements localization.

We intend to make profitable the Analytic knowledge based approach we used, by integrating into the system informations of different natures (prosodic, micro-prosodic, study of transitions etc.).

Moreover, in the context of this work, we have estimated the real interest of assessing the influence of temporal factors (ageing process of the voice) and psycho physiological factors upon the speaker's variability and specificity. These issues will be examined in our next research work.

BIBLIOGRAPHY

- [1] J.F. Bonastre, H. Méloni, *Etude de la variabilité spectrale pour la caractérisation du locuteur* ; 19^{èmes} JEP, Bruxelles, Belgique, 1992.
- [2] J.F. Bonastre, H. Méloni ; *Speaker Recognition and Analytic Process* ; Proc. 3rd European Conference on Speech Communication and Technology, 21-24 septembre, Berlin, Germany, 1993.
- [3] J.F. Bonastre ; *Stratégie Analytique Orientée Connaissances pour la Caractérisation et l'Identification du Locuteur* ; Thèse de l'université d'Avignon, 1994.
- [4] Calliope ; *La parole et son traitement automatique* ; Masson, 1989.
- [5] P. Durand ; *Variabilité acoustique et invariance en Français consonnes occlusives et voyelles* ; Edition du CNRS, 1985.
- [6] M. Rossi ; *De la quiddité des variables* ; Actes du Séminaire Variabilité et spécificité du locuteur, Marseille Luminy, 20-21 juin 1989, pp. 11-31, 1989.
- [7] J. Thompson, J. S. Mason ; *Within class optimization of cepstra for speaker recognition* ; Proc. 3rd European Conference on Speech Communication and Technology, 21-24 septembre, Berlin, Germany, pp. 165-168, 1993.
- [8] H. Van den Heuvel, T. Reitveld ; *Speaker Related Variability in Cepstral Representations of Dutch Speech Segments* ; ICSLP 92, Canada, Vol. 2, pp.1581-1584., 1993