STRUCTURAL (PHONETIC) EVALUATION OF DISSIMILARITIES
FUNCTIONS USED IN SPEECH RECOGNITION

H. YÉ 1 M.J. CARATY 2 L.J. BOÉ 1 D. TUFFELLI 1
Institut de la Communication Parléé
INPG/ENSEERG – Université Saintal UA CNRS n° 368
46 Avenue Félix Viallet F38031 Grenoble Cedex FRANCE
2 LAFORIA UA CNRS n° 1095 Laboratoire de Traitement de la Parole
Université Pierre et Marie Curie (Paris 6)
4 Place Jussieu F75232 Paris Cedex 06 FRANCE

ABSTRACT
We have evaluated 17 variants of 6 dissimilarities : PLOMP, Log Likelihood Ratio, Cepstrum, Mel Frequency Cepstrum Coefficients, Weighted Slope Metric, and Spectral Peaks Adjustment derived from FFT and/or LPC analysis with two types of integration (KLATT and ZWICKER). We used as "references" synthetic and natural vocalic stimuli for which we have a phonetic structural representation.

The intervocalic dissimilarities were used as input for a multidimensional analysis (KRUSKAL) to obtain an output space that we compared with the acoustic one from the data. The appraisal of the 2 spaces - the first one corresponding to F1-F2, derived from acoustic analysis and the second one rebuilt from dissimilarities as input of the Multidimensional Scaling KRUSKAL - allows us to compare dissimilarities and to make an extrinsic (phonetic) judgment on their behavior. We have used 5 criteria based on the capability of these processings to deliver vocalic dissimilarities that could be interpreted in terms of phonetic description (acoustic representation).

This comparative evaluation of dissimilarities can guide better choices regarding their application to automatic recognition and also in the domain of phonetic analysis, including perceptual simulation.

INTRODUCTION
The aim of this work is to evaluate, in the light of a priori knowledge, dissimilarities used in the frame of Speech Recognition.

We have evaluated 17 variants of 6 dissimilarities that are representative in the domain of automatic speech recognition and/or coding with integration or not of perceptual knowledge. Using as "references" synthetic vocalic stimuli and natural speech stimuli, for which we have a phonetic structural representation, our criteria were based on the capability of these processes to deliver vocalic dissimilarities that could be interpreted in terms of acoustic and perceptual phonetic knowledge.

The intervocalic dissimilarities were used as input for the KRUSKAL multidimensional analysis to obtain an output space that we compared with the acoustic one (F1-F2) plane from the data : the two first dimensions of the analysis being linked closely to F1 and F2. For this comparison, we have used 5 criteria of representation.

This work is a continuation and an expansion of two previous papers which served us as preliminary studies [1,2].

CHOICE OF STIMULI
To test the dissimilarities, we used synthetic vocalic stimuli of natural speech whose formant frequencies are given in table I. Synthetic stimuli were generated from reference area functions for the 11 French vowels [3] and also using a pole-zero synthesizer with Fo and intensity contours corresponding to natural speech [4]. The 11 Bedrod vowels were recorded by a male speaker in CV contexts and in carrier sentences ; formants were measured manually (cepstral method) by an expert phonetician [5]. The vocalic system of this language offers the advantage of presenting a very satisfactory exploitation of the maximal space, with three series: front, central and posterior vowels. Moreover, it presents two very close vowels [ə, e] which will allow a precise test of dissimilarities. The two reference systems are presented in fig. 1 and fig. 2 in the F1-F2 plane.

CHOICE OF DISSIMILARITIES
We retained 6 types of dissimilarities : PLOMP [6], Log Likelihood Ratio [7], Cepstrum [8], Mel Frequency Cepstrum Coefficients [9], Weighted Slope Metric [10] and Spectral Peaks Adjustment [11] some derived from FFT and/or LPC analysis, with two types of integration, KLATT [12] and ZWICKER [13] :

- LPC analysis + ZWICKER bandwidth integration = LPC ZWI
- FFT analysis + ZWICKER bandwidth integration = FFT ZWI
- LPC analysis + KLATT bandwidth integration = LPC KLA
- FFT analysis + KLATT bandwidth integration = FFT KLA
French synthetic stimuli
F1  F2  F3
l  270 2380 2970
o  390 2230 2650
e  530 1790 2470
a  690 1330 2580
å  680 1210 2480
y  270 1770 2300

Bedjond natural speech
F1  F2  F3
l  166 2033 2916
o  366 1883 2450
æ  550 1649 2300
u  290 716 2233
ø  400 1200 2233
æ  550 1366 2366
u  200 716 2233
ø  383 766 2266
ø  583 949 1783

Table I. Formant frequencies of synthetic (for French) and natural (for Bedjond) stimuli (Hz).

Finally we obtain a total of 17 dissimilarities. We calculated for the 11 French vowels and for each of these dissimilarities, "distances" between the 11 French synthetic stimuli taken two by two (a total of 17 matrices). For Bedjond, 4 dissimilarities were selected among those that gave best results, for representation and discrimination [2] and the same type of calculation was also applied.

SPACES OBTAINED FROM DISSIMILARITIES

Given these dissimilarities between vowels, we used a KRUSKAL multidimensional analysis [14], specially adapted for reconstruction of spaces. Fig. 3-6 present, e.g., spaces obtained from these dissimilarity matrices that were calculated by APS-LPC (fig. 3), MFCC-FFT-ZWICKER (fig. 4), LLR-LPC (fig. 5), PLOMP-FFT-ZWICKER (fig. 6). Results of these analyses were evaluated in the light of representation criteria.

---

**Fig. 1** French reference vowels in the F1-F2 plane.

**Fig. 2** Bedjond reference vowels in the F1-F2 plane.

**Fig. 3** Projection of APS-LPC.

**Fig. 4** Projection of MFCC-FFT-ZWICKER.

**Fig. 5** Projection of LLR-LPC.

**Fig. 6** Projection of PLOMP-FFT-ZWICKER.
REPRESENTATION CRITERIA

A global estimation of the fitness of these representations was calculated using 4 representation criteria:

(i) criterion C1 estimates the maximization of "distances" between the 3 extreme cardinal vowels [i, a, u];

(ii) criterion C2 reflects the goodness of the dichotomy between front [i, e, e] and back vowels [u, o, a];

(iii) criterion C3 evaluates the good positioning of the front rounded vowels [lj] for French, and central for Bedjond [i, a, o, e] in relation to the other series;

(iv) criterion C4 appreciates the position of vowels of intermediary quality [e, e, o, e, a] in relation to extreme vowels [i, y, e, u, a].

With these criteria, we can evaluate qualitatively the performances of our dissimilarities. Three values can be attributed per criterion: +1, 0, -1 corresponding respectively to a good, mediocre or bad representation. It should be made clear that we are dealing here with a global (total) estimation, which comprises some subjectivity.

Table II presents scores obtained by the 17 dissimilarities, classified in order of performances.

<table>
<thead>
<tr>
<th>Dissimilarity</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>APS-LPC</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+4</td>
</tr>
<tr>
<td>PLOMP-FFT-ZWI</td>
<td>+1</td>
<td>+1</td>
<td>0</td>
<td>+1</td>
<td>+3</td>
</tr>
<tr>
<td>PLOMP-LPC-KLA</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>0</td>
<td>+3</td>
</tr>
<tr>
<td>CEP-LPC-LIN</td>
<td>+1</td>
<td>+1</td>
<td>0</td>
<td>+1</td>
<td>+3</td>
</tr>
<tr>
<td>MFCC-FFT-ZWI</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>+2</td>
</tr>
<tr>
<td>MFCC-LPC-ZWI</td>
<td>0</td>
<td>+1</td>
<td>+1</td>
<td>0</td>
<td>+2</td>
</tr>
<tr>
<td>ITAKURA</td>
<td>+1</td>
<td>+1</td>
<td>0</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>MFCC-LPC-KLA</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>+1</td>
</tr>
<tr>
<td>CEP-LPC</td>
<td>+1</td>
<td>+1</td>
<td>0</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>WSM-FFT-ZWI</td>
<td>+1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>+1</td>
</tr>
<tr>
<td>PLOMP-LPC-ZWI</td>
<td>0</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>CEP-LPC-QUA</td>
<td>0</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>MFCC-FFT-KLA</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>WSM-FFT-KLA</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>WSM-LPC-KLA</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>WSM-LPC-ZWI</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

For natural Bedjond speech stimuli, we retained APS (fig. 7), PLOMP-LPC-ZICKER (fig. 8), MFCC-FFT-KLATT (fig. 9), LLR (fig. 10) which were proved to be efficient for discrimination in our previous study [2]. We used the same criteria C1 C2 C3 C4, and to estimate the separation of [a] and [o] we added fifth criterion C5.
CONCLUSIONS
The basic results are the following:
* Scores are clearly better with synthetic stimuli and this is not a surprise; note that we keep the same hierarchy.
* All processings do not constitute a maximal space having as limits the extreme vowels [i, a, u], whose reference role is fundamental in nearly all languages.
* The order of closed/open vowels is difficult to recover whatever the nature of the stimuli.
* Vowels of the front rounded and unrounded series for French are often confused and even in the case of Bedjond, which has central vowels, there are confusions with front or back vowels.
* The very close vowels [e] and [o] in the F1-F2 plane are differentiated by the all the processings using presumably upper formants F3 and F4.
* If one takes into account the set of results for synthetic stimuli and natural vowels, it is APS, MFCC-FFT-KLATT dissimilarities that best respect the phonetic space, followed by ITAKURA, PLOMP-LPC-ZWICKER dissimilarities which deliver acceptable results, these two latest being better in discrimination [2].

It is not our intention to select an "optimal" dissimilarity that could be used systematically. We think on the contrary - and results confirm it - that one should have a battery of processings, each of them presenting advantages for such and such a task. In a classificatory approach, we would be inclined to propose dissimilarity measures which "respect" best knowledge coming from phonetic typologies.

This work is only relevant to vocalic systems, for which certain aspects need to be explored: vocalic dissimilarity and dispersion, inter and intra-speaker dissimilarity and variability...

ACKNOWLEDGEMENTS We are very grateful to Dr. Rodolph Sock (from Institut de la Communication Parlée) for translating the text.

REFERENCES