PHONETIC UNIT LOCALIZATION IN A MULTI-EXPERT RECOGNITION SYSTEM

Hélène Tattegrain and Jean Caelen

Institut de la Communication Parlée INPG/ENSEERG - Université Stendhal UA n°368 46, av Félix Viallet 38031 Grenoble cedex FRANCE

ABSTRACT

This paper describes an acoustic-to-phonetic decoder (APD) based on a mixed strategy: a) bottom-up which hypothesizes the most robust information about the speech signal, b) top-down which makes some verifications about the acoustic features or about the macro-class localization on the speech signal. In this paper, only the bottom-up strategy is described.

In our system, a phoneme is described as a phonetic network whose nodes are mapped onto the acoustic signal. The coarse phonetic description then uses five phonetic networks whose nodes correspond to the acoustic phases of the analyzed sound in the speech signal. These phases are extracted by automatic segmentation using different parameters (energy, pitch, formant frequencies, acoustic cues from an ear model).

The group APD is divided into three steps: a) the first step localizes pseudo-phonetic segments (called acoustic phases) on the signal and defines phoneme boundaries according to a macro-class description (stop consonants, fricatives, other consonants, vowels and pauses); b) context-sensitive rules are then applied in order to filter out the most improbable solutions; c) the third step labels the most significant phase of each phoneme by acoustic features (using Bayesian methods).

In this paper, the performance is measured by the comparison between labels generated automatically and labels generated normally: for example, detection of plosive burst rates 97% while detection of occlusive phonetic network rates 94.3%. This strategy is written in Prolog II.

INTRODUCTION

The role of the APD (Acoustic-Phonetic Decoder) still a very difficult problem for two major reasons due to:

- the APD projection the little acoustic observation space into the large phonetic pattern space.

- the wide variability of the acoustic observations.

The breakdown between the acoustic substance and the phonetic pattern spaces prevent the APD to be considered as successive transformations from continuous to discontinuous representations: the two structures preexist and the APD maps out the relations between the acoustic micro-structure of the signal and the implicit phonetic macro-structure. So that the essential work is to describe suitably the phonetic macro-structures which are hidden behind the general knowledge of phonetic sciences. Then the mapping between the two structures can be made by the matching between the two models which represent the above two structures (macro and micro-structures).

We indicate that the models represented by HMM (Hidden Markov Models) have more clear perspectives than the methods which are derived from the ES (Expert Systems) techniques in which the phonetic structures are included in production rules whose organization is very dependent on the used strategy. In the other hand the disadvantages of the HMM models are:

- it do not provide us explicable models because of the blind learning of these models.

- it prevent several sources of knowledge (perceptive, articulated) to be considered.

The two approaches (ES, HMM) are compatible if the notions of "phonetic models of references" and the "matching" are clarified.

In the ES techniques we are led to define the models, which result from the phonetic macro-structure, under networks from. In these networks the transitions between the states are formulated by knowledge and appropriate rules which manipulate this knowledge. The matching identify the phonetic macro-structures over the micro-structure of the signal (or inversely) by a convenient searching over all the networks.

The ES techniques have an advantage that they separate between the knowledge of the two structures:

- the acoustic knowledge which is very connected to the parameterization procedures (cues, correlation, etc).

- the phonetic knowledge which is essentially conveyed by the features (symbolic information which are taken in a very large view).

Then the matching can be considered as a general process which is independent of the used knowledge.

The APD can either predicate the phonetic macro-structures from the micro-structures or verify the presence of phonetic macro-structures onto the micro-structures of the signal. Figure 1 illustrates the process of acoustic phonetic decoding.

This point of view of matching allows us to standardize the recognition procedure by overtaking the notions of bottom-up or top-down strategies.

Certainly if the phonetic units are known a priori by the decoder then the matching can be considered as a top-down verification or bottom-up research in the other case, but the mechanism of inference is still the same: only the contents of the phonetic networks are to be modified for weighting selectively the macro-phenomena and the micro-phenomena. It is the point of the difference between the AI (Artificial Intelligence) methods and the stochastic methods: the details which are very important in phonetics can be taken into account more easily by the rules bases than by the probability coefficients which smooth the rare phenomena or the weak amplitude.

The APD in the DIRA system (Integrated oral Dialogue and Automatic Recognition) is guided by the supervisor [6]. It can work in two modes:

a. Proposition mode: it relieves robust information in the form of macro-features, and after filtering and refining it relieves a list of features according to the context.

b. Verification mode: (b1) It verifies by "Spotting" the presence or the absence of an information in demand by the supervisor--this information can be of diverse nature like features, macro-features, prosodic markers...etc.(b2) It returns to Rendezvous point which is fixed by the supervisor (The Rendezvous point is a land back point for which the supervisor returns after a failure in the recognition process).
A KNOWLEDGE REPRESENTATION

A.1. MACRO-CLASSES

French sounds can be divided into macro-classes such as: stop consonants, voiced occlusives, voiced fricatives, voiceless fricatives, nasal consonants, glide consonants, semi-vowels, vowels and pauses.

Some sounds have multiple realizations. For example, the semi-vowels can inherit features from vowel as well as consonant, \(/v/ is analyzed both as a voiced fricative and as a voiced consonant. Some macro-classes have been grouped in order to increase the rate of correct recognition of the robust consonant, /v/ is analyzed both as a voiced fricative and as a voiced consonant.

Finally, there are five macro-classes: stop consonants, fricatives (voiced and voiceless), other consonants (voiced, nasal, glide), vowels and pauses. Each macro-class corresponds to a phonetic network which represents knowledge of the phonetic macro-structure.

A.2. PHONETIC NETWORK

A phonetic network \( R_j \) is defined by 5-tuplet:

\[
R_j = ( j, S(j), T, S_0(j), \pi_j )
\]

with \( j \) : network name, \( S(j) \) : phonetic nodes set, \( T \) : transition set \( t_i, S_{t_i} \) : initial node, \( S_0(j) \) : final node.

\( S(j) \) represents all possible realizations of acoustic phases for a given macro-class. For example:

\[
S(\text{fricatives}) = \{ s_k, s_p, s_t, C_j, A_j \}
\]

This figure shows the diversity of a fricative realization. This figure shows the diversity of a fricative realization. The following is an example of a transition rule. It must be realized in order to reach the "vocalic friction" node.

\[
\text{If } (\text{cues}) \text{ or } (\text{previous state} = "\text{vocalic friction}"
\]

We see in this rule the three types of constraint and the actions we have described.

A.3. FRICATIVE NETWORK

The figure 2 gives the different possibilities to connect these phases.

![Figure 2: Fricative Network](image)

Fig 2 : Fricative Network. \( S_j, Z_j, S_2 \) correspond to the transition rules. The states are in circles (except for beginning and end nodes). Procedural actions are in rectangles.

We describe the fricative network in order to explain the mechanism of the rules. This network has seven nodes (five acoustic phases and two factious states):

- beginning : (network entry)
- vocalic friction : (beginning of friction after a vowel or a voiced consonant)
- closure onset : (closure before friction)
- closure offset : (closure after friction)
- voiceless friction : (friction without voiced sound)
- voiced friction : (friction with voiced sound)
- end : (network exit)

This figure shows the diversity of a fricative realization. The following is an example of a transition rule. It must be realized in order to reach the "vocalic friction" node.

\[
\text{SZ1 "Vocalic Friction"}
\]

B. RECOGNITION STRATEGY

B.1. GENERAL PRESENTATION

![Figure 3: General Architecture of Acoustico-Phonetic Decoder](image)

Fig 3 : General Architecture of Acoustico-Phonetic Decoder.

The APD architecture have the following modules:

- A module of acoustic analysis which generates the signal parameters: 24 channel spectrum using an ear model [4], 6 cues (acute/grave, open/close, diffuse/compact, sharp/flat, mat/strident, continuous/discontinuous) [5]. It segments the signal using the cues and energy variations [11]. It computes the pitch using the AMDF method followed by dynamic programming [1], and the formants using Markov's method [2].
- A decoder module is built in two parts. (a) The bottom up APD which provides some robust information with or without indication from the supervisor. (b) The top down APD which verifies the sound over the signal in demand by the supervisor.

The bottom up APD is built in three parts.

The first deals with the localization of the macro-classes with the help of the networks. The context constraints are
supposed realized.

The second is the filtering of candidates. After the first part we get a set of macro-class hypotheses. We know the possible contexts around the candidate so we can cancel some improbable solutions and systematic errors (cf results).

The third part is the accurate recognition of acoustic features based on the use of the most relevant acoustic phases to distinguish the different sounds in the same macro-class.

B.2. BOTTOM UP APD RESULTS

The principal quality of the bottom up APD is its robustness (its capacity to give some robust information even if it is not enough for a fine recognition). This robustness depends on the performance of localization: it is very easy indeed to rectify an acoustic feature but it is very difficult to reject a bad localization, as no new information is available on the speech signal.

![Speech signal sample](image)

**Fig 4**: A speech signal sample from a male speaker: "I got to go" and results of "consonants" network (line 1), of "fricatives" network (line 2), of "stop consonants" network (line 3), of "pauses" network (line 4) and of "vowels" network (line 5).

In this figure, we present only here, the results about the localization step. At the beginning of each sentence, the network "pauses" and "stop consonant" have been kept: three SI (silence), one DP (speech beginning by a pause), three OT (total occlusion), and one BF (fricative burst for the stop consonant). The filtering phase will suppress the stop consonant solution because it is at the beginning of a sentence and the burst isn't significant in this case.

B.2.1 ROBUSTNESS OF DETECTION

Using the speech signal file (and manual labeling) of BDSON (French Sounds Data Base), the duration of each phoneme is known. With \( Z \) segment of candidate and \( Z^* \) segment of referred label, we define an evaluation parameter RR (Recovering Rate) for the macro-class localization regarding to the previous segmentation:

\[
RR = \frac{\text{length}(Z \cap Z^*)}{\min(\text{length}(Z), \text{length}(Z^*))}
\]

A localization is correct if \( RR \geq 50\% \) and the detection robustness score \( R \) is arrived at simply by adding the good solutions (score D) and the holes.

The principal errors are due to the glide consonants because of their contextual variability, and to the nasal consonants, especially where some like /m/ have a high level of energy like a vowel, through /n/ gives good results.

<table>
<thead>
<tr>
<th>PHONEMES</th>
<th>MALE</th>
<th>FEMALE</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOP CONSONANTS</td>
<td>D Hole</td>
<td>R D Hole</td>
</tr>
<tr>
<td>STOP CONSONANTS</td>
<td>92.46% 0.69%</td>
<td>93.15% 94.59%</td>
</tr>
<tr>
<td>VOICELESS FRICATIVES</td>
<td>97.87%</td>
<td>97.87% 90.00%</td>
</tr>
<tr>
<td>VOICELESS FRICATIVES</td>
<td>88.33% 88.33%</td>
<td>95.01% 95.91%</td>
</tr>
<tr>
<td>NASAL CONSONANTS</td>
<td>84.81% 84.41%</td>
<td>82.25% 80.67%</td>
</tr>
<tr>
<td>LIQUID CONSONANTS</td>
<td>63.21% 5.75%</td>
<td>68.96% 44.77%</td>
</tr>
<tr>
<td>VOICELESS OCCLUSIVES</td>
<td>92.18% 92.18%</td>
<td>93.75% 4.16%</td>
</tr>
<tr>
<td>SEMI-VOWELS</td>
<td>97.43% 97.63%</td>
<td>97.03% 2.55%</td>
</tr>
<tr>
<td>VOWELS</td>
<td>93.64% 1.15%</td>
<td>94.72% 95.44%</td>
</tr>
<tr>
<td>PAUSES</td>
<td>100% 0%</td>
<td>100% 98.75%</td>
</tr>
</tbody>
</table>

**Tab 1**: Detection Robustness score \( R \) after the localization phase, in 50 sentences for a man and 40 for a woman.

In conclusion, these results show good localization of acoustic phases.

\( R_{\text{Moy}} = 90.73\% \) on 1275 sounds for a male speaker and \( 93.56\% \) without the glide consonants.

\( R_{\text{Moy}} = 91.40\% \) on 1005 sounds for a female speaker and \( 96.30\% \) without the glide consonants.

B.2.2. FINESS OF DETECTION

Another evaluation method consists in verifying that each labeled phoneme corresponds to a candidate of the same macro-class with a rate \( R' \geq 50\% \). With \( Z \) segment of candidate and \( Z^* \) segment of referred label, we define an evaluation parameter \( R' \).

\[
R' = \frac{\text{length}(Z \cap Z^*)}{\text{length}(Z^*)}
\]

The Rate "FI", test of detection fineness can be computed by the addition of referred phonemes which have a candidate in the same place such as \( R' > 50\% \).

<table>
<thead>
<tr>
<th>PHONEMES</th>
<th>STOP CONSONANTS</th>
<th>VOICELESS FRICATIVES</th>
<th>VOICELESS FRICATIVES</th>
<th>NASAL CONSONANTS</th>
<th>VOWELS</th>
<th>OTHER CONSONANTS</th>
<th>WOLES</th>
<th>PAUSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOP CONSONANTS</td>
<td>133 0 1 2 10 0 0 1 0 90.67%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOICELESS FRICATIVES</td>
<td>0 90 5 0 8 0 1 0 0 96.59%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOICELESS FRICATIVES</td>
<td>1 0 3 0 2 0 0 0 0 50.30%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NASAL CONSONANTS</td>
<td>4 0 45 67 89 55 14 43 0 94.59%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOWELS</td>
<td>1 0 1 10 40 4 23 441 0 89.22%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTHER CONSONANTS</td>
<td>0 0 0 0 0 2 1 0 0 100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAUSES</td>
<td>2 0 0 0 0 0 0 0 0 100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Tab 2**: Confusion Matrix. FI, rate of fineness of detection, individual network reliability, with the 50 male sentences.

Some systematic errors appear as follows:
- the network "stop consonants" make a mistake on pause at the beginning of sentences or in the context of a glide consonant.
The network "fricatives" may detect phonemes which are near a fricative.
- voiceless considered by the network to be consonants are very often closed vowels.
- consonants considered by the network to be vowels are very often nasal or glide consonants.
- the stop consonants which were not detected are viewed as by the fricatives (because of the burst) or as consonants (because of the post-vocalic implosion) or as pauses (with a very short silence phase).
- The female speaker confusion matrix presents the same type of errors.

An alternative evaluation, more constrained consist of the examination of the results from a negative view point and generates the recovering matrix. In this evaluation, we add all the candidates with the value RR' as in the confusion matrix.

The principal problem is at the beginning of the sentences (but it consonants, /d/, /r/ after stop consonants) and 12 pauses.
- the network stops consonants because the burst is long and fricative (this error can be suppressed by the filtering phase) and for the glide near occlusives (this problem could be examined with phonotactic rules in the top-down APD).
- the network "voiceless fricatives" detects 13 stop consonants because the burst is long and fricative (this error can be suppressed by the filtering phase). Considering the other errors, /v/ (1 final 1 voiceless), 1 /a/ (before one /s/), 1 /l/ (before /l/), 1 /i/ (after /i/), 7 /R/ and 1 /l/, their case must be examined at the phonotactic rule level in the bottom up APD.
- the network "voiceless fricatives" detects 7 stop consonants for the same reason as for the "voiceless fricatives", 1 /a/ (before 1 /s/) 2 /R/, 1 /l/ (unvoiced). We do not have a large number of candidates because a lot of voiceless fricatives are detected by the "consonants" network (a lot of /l/ are vocable sound).
- the network "consonants" detects 10 stop consonants because the onset before the closure is viewed as a consonant closure. This error can be easily suppressed by the filtering phase.
- the network "vowels" detects 5 /l/, 3 /i/, 1 /a/, 29 /R/, 29 /l/, 17 /m/, 8 /u/, 6 /i/. The principal problem is due to the glide and nasal consonants.
- the "pauses" network detects 3 /R/, 1 /l/, 1 /l/ (this error could be suppressed by prosody module rules [10]) and 88 stop consonants (suppressed by filtering phase).
- the "neutral sound" network detects 4 voiceless fricatives, 17 consonants and 4 vowels. It correctly detects the sounds of low energy.
- The female speaker results are very similar. The only difference is with the voiceless fricatives /s/ which are considered as "stop consonants" by the relevant network because their energy is low.

This test set is made without taking into account the hypotheses on the context in order to suppress the consequences of errors in previous localization or the consequences of false linguistic hypotheses. That is why, we can offer an objective rate without improving them by the use of the constraints, favorable in the most of cases, induced by context. In the normal running APD is guided by the supervisor which optimizes the APD probability in deciding the best strategy, (bottom up or top down).

**CONCLUSION**

The starting point for this work is the knowledge base proposed by the expert phonetician. The system has been written in Prolog language using the phonetic and acoustic knowledge.

The unit localization results of bottom up APD shows the knowledge base robustness. Phonetic networks provide an adequate sound representation. The errors are due essentially to glide consonants (50%) which are considerably influenced by the context and to nasal consonants (30%) which have a high level of energy (and in fact are very close to vowels). These errors are suppressed by the step which follows using the context together with phonotactic rules.

The knowledge structure of phonetic networks can separate clearly the phonetic macro-structure and acoustic micro-structure of the speech signal (on the paradigmatic and syntagmatic axes). The concept of phonetic network produces a connection between the stochastic model (HMM) and the classic rule based system in which the speech macro-structure is not sufficiently visible. This notion also allows us to work in two modes (a "proposition" mode, a "verification" mode) without modifying the procedures which make the correspondence between the acoustic micro-structure and the phonetic macro-structure. This concept can also offer the possibility for the implementation of a parallel strategy on specialized computers.

**REFERENCES**


**Tab 3**: Recovering Matrix. F2, rate of detection fineness, global network reliability, with the 50 male speaker sentences.

<table>
<thead>
<tr>
<th>NETWORK</th>
<th>STOP CONSONANTS</th>
<th>VOICELESS FRICATIVES</th>
<th>VOICE FRICATIVES</th>
<th>NASAL CONSONANTS</th>
<th>OCCASIONAL CONSONANTS</th>
<th>VOWELS</th>
<th>MUSCLE NOISES</th>
<th>PAUSES</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOP CONSONANTS</td>
<td>133</td>
<td>0</td>
<td>2</td>
<td>21</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>77</td>
<td>56.7%</td>
</tr>
<tr>
<td>VOICELESS FRICATIVES</td>
<td>17</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>70.0%</td>
</tr>
<tr>
<td>VOICE FRICATIVES</td>
<td>15</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>11.7%</td>
</tr>
<tr>
<td>NASAL CONSONANTS</td>
<td>15</td>
<td>0</td>
<td>44</td>
<td>47</td>
<td>89</td>
<td>76</td>
<td>21</td>
<td>119</td>
<td>47.7%</td>
</tr>
<tr>
<td>OCCASIONAL CONSONANTS</td>
<td>2</td>
<td>0</td>
<td>14</td>
<td>25</td>
<td>42</td>
<td>13</td>
<td>23</td>
<td>441</td>
<td>1</td>
</tr>
<tr>
<td>VOWELS</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>94.9%</td>
</tr>
<tr>
<td>MUSCLE NOISES</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>94.9%</td>
</tr>
<tr>
<td>PAUSES</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>

**Actes 14ème Journee d'Etudes de la Parole, Paris, France, September 1989**

**EUROSP**

**2259**