SPEAKER-INDEPENDENT SPEECH-RECOGNITION USING ALLOPHONES

E. Barkowsky
Institut de la Communication Parlée
Institut National Polytechnique de Grenoble
46, Avenue Félix Viallet
38000 GRENoble, FRANCE.

D. Jouvet
Centre National d'Études des Télécommunications
Route de Trégastel
22300 LANNOY, FRANCE.

ABSTRACT

This study concerns the determination of the allophones that are necessary for achieving a good recognition of the French numbers by a speech recognition system based on a Markov modelling approach. The allophones have been distinguished, for the vowels, by the transition variations at the "onset" and at the "offset", and for the consonants, by their phonetic characteristics.

For this specific application, using an average of 2 allophones by phoneme and a few "clusters", we achieved 94.9% correct recognition rate on the adult numbers, for 13 speakers that were not in the training set.

INTRODUCTION

A speaker-independent speech-recognition system has to deal with all the possible acoustical realizations of the words in the vocabulary. The variations result from individual differences, the acoustic realization of the context, and articulatory effects. The recognition system, based on a Markov modelling approach, can handle part of these variations through the automatic training procedure. However, the basic units, used to describe the words (usually phonemes), have different acoustical realizations depending on the context. If one uses a specific acoustic model for each phoneme in each context, the total number of basic units would probably be too high.

But, for any phoneme, several contexts may have nearly the same influence on its acoustical realization. So a good trade-off, between accuracy and complexity, is to use different acoustic models, for a given phoneme, depending on the context in which it occurs.

This is why we are studying the allophones, i.e. the voice variants corresponding to a particular acoustical realization. It is worthwhile mentioning that this study concerns only a specific application, namely the French numbers between 0 and 999, and thus has no pretention to be a full study of the French allophones. Nevertheless, the set of allophones determined in this study may be extended as needed to fit a new vocabulary. The French numbers consist of nearly the same influence on its acoustical realization. So a good trade-off, between accuracy and complexity, is to use different acoustic models, for a given phoneme, depending on the context in which it occurs.

This study, concerning the determination of the allophones, was conducted using the spectrogram of the signal in association with the pitch and the waveform.

This study, concerning the determination of the allophones, was conducted using the spectrogram of the signal in association with the pitch and the waveform.

VOWELS ALLOPHONES

One of the main acoustical realizations of the context influence on the vowel is the transition of the formants of the "onset" or of the "offset". For practical reasons we grouped together the apico-dental and the pre dorso-alveolar contexts because the transition of the formants they induce are very similar [3].

Consonant Influence

From the loci theory [2, 3], which explains the transition of the formants at the "onset" or of the "offset" of the vowels by the point of articulation of the adjacent consonant, we defined classes for the consonants. We grouped together the apico-dental and the pre dorso-alveolar contexts because the transition of the formants they induce are very similar [3].
independent of this influence. For example, the /a/ appearing in the third syllable of a 6 syllables sense-group adjacent to a pause. Some studies [10] note an increase of the duration of the nasals in the context of a sense-group adjacent to a pause. For some speakers having a rather slow speaking rate, the nasals may occur, for some speakers, even in an intervo-

**NEUTRAL VOWEL — SCHWA**

The neutral vowel should be treated like a possible occurrence place rather than an acoustical realization pattern. Theoretically, in French, at a slow speaking rate in a careful articulation manner, it is possible to pronounce a schwa at the end of every isolated word ending by a consonant. However, for connected words such as the numbers, it is possible to pronounce a schwa at the end of every isolated word ending by a consonant. After that, we treat the case of the epenthetic schwa from the vowels /oe/ and /eu/. For some speakers, even in an intervo-

**CONSONANTS ALLOPHONES**

The different realizations of the consonants, are first described using phonetic characteristics such as duration and place of articulation, as modifiers applied to the "standard" realization. An initial voiced consonant may also have a very short duration, and even vanish, in which case the only remaining cues are the formant transi-

**Voicing feature**

*Vocalization r, v. Vowel —— Vowel
*Labialization  z. /ou/,/on/ —— *
*Palatalization{ d. * —— /i/
*Fricatization: The unvoiced realization of /r/ is usually due to an adjacent voiceless consonant, which occurs for the initial position of a sense—group and for positions at the "onset" of the following vowel (for example /y/ in /y.i.t/ or /v/ in /v.in/). For these reasons we have to define, in an initial position, just after a pause. 2 allophones with different acoustical realizations and different durations for the fricative /f/ and the nasalized /y/, one corresponding to a "standard" pronunci-

**Segmental Duration in French**

Herman, 1984.

**RECOGNITION TESTS**

The reference point, for measuring the improvement due to the allophones, is a phonetic based model, in which the words are described as sequences of phonemes, each of them being represented by the same acoustical model, independent of the context. However, because of the coarticulations effects, the sequences /i/./t/, /a/ and /y/ were considered as basic units and thus were represented by a single acoustical model. Using this description, we achieved 93.1 current classification rate on the whole numbers for the testing set. Using an average of 2 allophones by phonemes, introducing specific models to handle transitions between adjacent vowels, and keeping the 3 clusters mentioned above, we achieved 94.5% correct classification rate on the same data base, thus reducing the error rate by 25%.

**CONCLUSION**

This paper shows that a good description of the acoustic model changes the performances of a speech recognition system. As the coarticulations and the different pronunciations are predicted, the acoustical models have just to take into account the variations due to the different speakers. However, it needs to correctly predict all the coarticulations, it would be necessary to consider the individual specific application, the individual variations of the speakers and also the speech rate of the current sense-group. The set of allo-

**BIBLIOGRAPHY**


[4] M. Rossi, Y. Nishinuma, G. Mercier: "Indices acoustiques multiloueurs et indépendance du contexte pour la reconnaissance automatique de la parole"; Speech communication, North-


[1] D. Jouve, J. Monod, D. Dubois: "Un nouveau système de segmentation de la parole"; Les représentations en phonolo-

It seems there is a mix-up with the page numbers and content. The documents on pages 85.4.3 and 85.4.4 contain overlapping text. If you need specific content, please specify the page numbers or parts you are interested in.