SPEAKER-INDEPENDENT SPEECH-RECOGNITION USING ALLOPHONES

K. Bartkova
Institut de la Communication Parlée
Institut National Polytechnique de Grenoble
46, Avenue Félix Viallet
38000 GRENOBLE, FRANCE.

D. Jouvet
Centre National d'Etudes des Télécommunications
Route de Trégastel
22300 LANNION, 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 formant transitions at the "onset" and at the "offset", and for the consonants, by their phonetical characterization.

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 whole 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 various speakers, different possible pronunciations and coarticulation effects. The recognition system we used [1], 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 acoustical model for each phoneme in each context, the total number of necessary models would be pretty large. But, for any phoneme, several contexts may have nearly the same influence on its acoustical realization. So a good tradeoff, between accuracy and complexity, is to use different acoustical models, for a given phoneme, only when the context influence is different enough.

That is the reasons why we are studying the allophones as each of them corresponds to a particular acoustical realization. It is worthvile 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 theoretical research 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 use nearly all the vowels and half of the consonants of the French language. The strict syntax and the limited vocabulary restrict the number of contexts for each phoneme, thus, we were able to conduct a full study of the different contexts for this specific application.

After a description of the data base, this paper details the different realizations of the phonemes. For the vowels, we used mainly the transition of the formants, and for the consonants, their phonetical characterizations. We end the paper by an application of the allophones in a speech recognition system.

DATA BASE

The data base contains about 3000 French numbers between 0 and 999. They were recorded from 34 adult speakers (22 men and 12 women). All the speakers have a "standard" pronunciation, except one having a strong regional accent.

The table lists the different phonemes of the data base. For typographical reasons we denote the phonemes by one or two ascii characters, and we specified here the standard phonetical meaning when different from the notation used.

Vowels	Oral	i, ei(e), ai(ɛ), a, o(ɔ), au(o), ou(u), eu(ø), oe(œ), e(ə).
	Nasal	an(ɑ̃), in(ɛ̃), un(œ̃), on(ɔ̃)
Consonants	Plosive	d, t, k.
	Fricative	v, z, f, s.
	Liquid	r.
	Nasal	n.
	Semivowel	w, y(ɥ).

This study, concerning the determination of the allophones, was conducted using the spectrograms of the data 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 at the "onset" and at the "offset". For practical reasons, related to the implementation of the speech recognition system, we will treat separately the consonantic influence, the pause influence, the possible devoicing and the case of adjacent vowels.

Consonantic Influence

From the locus theory [2, 3], which explains the transition of the formants at the "onset" or the "offset" of the vowels by the point of articulation of the adjacent consonant, we defined 6 classes for the consonants. We grouped together the apico-dental and the predorso-alveolar contexts because the transition of the formants they induce are very similar [3].

Labial	f, v	(labio-dental)
Dental	t, d, n	(apico-dental)
	s, z	(predorso-alveolar)
Velar	k	
Labio-palatal	y	
Labio-velar	w	
Uvular	r	

Instead of measuring by degrees the displacement of the articulation point, or the aperture [4], during the realization of the vowel, we will characterize the allophones by their full context. The next table reports the vowels of the data base and the contexts in which they occur. Each row corresponds to a left context, and each column to a right context. "Lpal" stands for labio-palatal, and "Lvel" for labio-velar. In order to represent all the positions we add the "pause" and "vowel" contexts. They will be treated later on.

Left \ Right	Labi	Dental	Lpal	Vela	Uvul	Paus	Vowe
Labial		in	in	in		in	in
Dental	an	ai,an,eu	an	an	ei	an	an
	oe	i,in,ou		in	o	eu	ei
Lpal.		i					
Lvel.		a				a	
Velar		a,an,in			a		
Uvular	au	ai, au	au	au		au	au
	oe	an					
Pause		on				un	
Vowel		on				un	

Taking each vowel in each possible context would define a full set of allophones for this application. However, one can also define a subset by grouping together for each vowel some contexts which have nearly the same influence. We should point out that this grouping is different from one vowel to another. For example, for the oral vowel /au/ we can put together the right contexts "velar", "labial" and "pause"; but for the nasal vowel /an/ we will have to keep separate the realizations corresponding to the "velar" and "labio-palatal" contexts which induce important transitions on the formants.

Pause Influence

In general 3 different realizations are possible for the "onset" or the "offset" of a vowel when adjacent to a pause. Just after a pause, we can have a glottal stop, a synchronized or an aspirated (devoiced) beginning. Just before a pause, we can have a glottal stop, a synchronized or a devoiced ending. A synchronized beginning or ending corresponds to a progressive rising or falling of the pitch and of the intensity showing a synchronization between the vocal cords vibrations and the velum and the articulators movements. A devoiced beginning or ending results from a partial forward or backward assimilation, the pause having the same effect as a voiceless context.

Vowel devoicing

The voicing feature appears to be rather robust for the vowels. Only one context was strong enough in our data base to devoice a whole vowel. This context was /s.w_s/ for the vowel /a/ in the word "60" (/s.w.a.s.an.t/). After a devoicing of the /w/, the vowel /a/, surrounded by voiceless consonants loses its voicing feature and becomes coarticulated with the surrounding noise.

Adjacent vowels

For 2 adjacent vowels, belonging to different words, we noticed the following realizations: For **unstressed vowels** the transition of the formants is smooth and uninterrupted. For **stressed vowels** either a short pause (50 to 200 ms) appears between the vowels and they may start or end by a glottal stop, or the transition is realized by a glottalized vocalic portion having a low pitch.

Summary

Because of the implementation in the speech recognition system we group together the "pause" and the "vowel" contexts. The pause does not induce formant transitions, and the transitions between adjacent vowels are handled by specific acoustical models. In order to obtain a good representation of the various transitions of the formants we had to define an average of 2 allophones by vowel. The number of allophones used for each vowel, reported in the following table, does not take into account the pause influence and the possible devoicing.

Vowel	i	ei	ai	a	o	au	ou	oe	eu	an	on	in	un
Alloph.	2	2	2	3	1	3	1	2	1	4	1	4	1

SUPRASEGMENTAL INFORMATION

As this speech recognition system does not use pitch information, and also because for such short sentences the pitch is not a useful syntaxic hierarchical cue, the only suprasegmental information we have studied is the segmental duration. The importance of the vocalic duration is justified by the facts that, besides an obvious correlation between word and phoneme durations, the degree of perturbation of the formants by the context is strongly related to the length of the vowel. Also, the knowledge of the minimal duration is useful for designing the acoustical models. For these reasons we have started a statistical analysis of the segment durations.

The vowels before a pause are longer than the same in other positions. This agree with the fact that, in French, the stress is on the last syllable of the sense-group [5,6], which often corresponds for our application to the whole numbers. One of the most important acoustical realizations of the stressed syllables is the longer duration of the vocalic nucleus. In a closed syllable, the influence of the following consonant on the vocalic duration agree with previous studies [7, 8].

The vowel duration in a non final syllable, therefore unstressed position, is strongly correlated with the duration of the sense-group. However the duration of a stressed vowel is independent of this influence. For example, the /a/ appearing in the third syllable of a 6 syllables sense-group lasts 54 ms as regards to 82 ms when being in the unstressed syllable of a 2 syllables word. But, when the /a/ appears in a stressed syllable, it lasts 164 ms in a 6 syllables group even followed by a shortening consonant such as /t/ in French, compared to 144 ms, ceteris paribus, in a 2 syllables word.

NEUTRAL VOWEL - SCHWA

The neutral vowel should be treated like a possible occurence 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, this neutral vowel may be pronounced at the end of each of the individual words. For example, whether the schwa (e) is pronounced or not, implies 4 different theoretical patterns for a sequence like "55" (/s.in.k.an.t.(e).s.in.k.(e)/).

For a correct identification of the neutral vowel, one needs to use suprasegmental information such as the vocalic duration. The duration seems to be the more appropriate cue for differentiating the schwa from the vowels /oe/ and /eu/. For example, the duration of the schwa before a pause was always very short compared to the duration of the previous stressed vowel.

CONSONANTS ALLOPHONES

The different realizations of the consonants are first described using phonetical characteristics such as nasalization, labialization, etc, as modifiers applied to the "standard" realization. After that, we treat the case of the epenthetic sounds and the voicing feature.

Allophonic characterization

Nasalization: This concerns the stop consonants after a nasal vowel. The voiced stop /d/ may, by a forward coarticulation effect, become partly or completely nasalized. For the voiceless stops, a nasal consonant may be realized before it or even replace it.

Palatalization: This concerns the stop consonants in a right labio-palatal context, or before a palatal, anterior vowel.

Labialization: This concerns the fricatives followed by a labio-velar semivowel, or preceeded by a rounded posterior vowel.

Vocalization: This concerns the voiced fricative /v/ and the liquid /r/ in some intervocalic positions (for example /oe.v.in/ in "80" and /a.r.an/ in "40").

Fricatization: The unvoiced realization of /r/ is in a strict sense a fricative [9]. The devoicing, usually due to an adjacent voiceless consonant, may occur, for some speakers, even in an intervocalic context.

Rolled: This concerns, in our data, only the unvoiced /r/ after the voiceless stop consonant /t/. This realization is produced by a flapping (quasi occlusion) between the back of the tongue and the "velo-uvular" region.

Tense: For these data, the consonant duration vary a lot in two positions: first or last consonant of a sense-group adjacent to a pause. Some studies [10] note an increase in the tension of the articulators, the vocal cords and the velum for the initial position of a sense-group and for the stressed syllable of the group. We will denote as "tense" the corresponding realization of the consonants. This characteristic does not correspond to the feature "tense" as defined in some classical theories of segmental phonology [11], but rather defines some consonantic realizations appearing in specific contexts.

An initial voiced consonant may also have a very short duration, and even vanish, in which case the only remaining cues are the formant transitions 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 /v/ and the semivowel /y/, one corresponding to a "standard" pronunciation and the other to the "tense" realization. At the end of a sense-group, in a stressed syllable, the VOT of the unvoiced stops, when followed immediatedly by a pause have the same realizations as "tense" consonants for some speakers (important high frequency noise and a longer VOT).

Speaking rate and epenthetic sounds

For some speakers having a rather slow speaking rate we notice the realization of 2 epenthetic sounds, one consonantic and one vocalic. An unvoiced consonantic "closure" is realized in a context where the nasal consonant is preceeded by an unvoiced consonant; this occurs for the consonant /n/ preceeded by the voiceless stop /t/ or the devoiced fricative /z/. A neutral vowel (schwa) may occur when the voiced realization of /r/ is followed by a voiced consonant.

Voicing feature

The voicing feature, for the consonants, is often inaccurate and is strongly influenced by the context. In fact, its modification, due to coarticulations effects, appears to be the same for most of the consonants: stops, fricatives and semi-vowels. For voiced consonants, the pause has the same influence as an unvoiced context, and thus implies a partial or total devoicing by a forward or backward assimilation. The following table gives, for the specified contexts the consonants for which the voicing feature may be modified:

Consonants	Left context	Right context
d, z, v.	Pause	Vowel
z.	Vowel	Pause
y, w.	Unvoiced consonant	Vowel
t, k.	Vowel	Vowel

Summary

The following table lists for each characteristics the consonants that are affected, and the contexts that induce this modification by a forward or backward assimilation. The "*" denotes an irrelevant context (ie anything).

Characteristics	Consonants	Contexts
Nasalization	t, d, k.	/an/,/in/ — *
Palatalization	t.	* — /y/
	d.	* — /i/
Labialization	s.	* — /w/
	z.	/ou/,/on/ — *
Vocalization	r, v.	Vowel — Vowel
"Rolled"	r.	/t/ — *
Fricatization	r.	/t/ — *
	r.	Vowel — Vowel
Tense	v, y.	Pause — *
	t, k (VOT)	* — Pause

RECOGNITION TESTS

We applied this study to the speaker independent recognition of the French numbers between 0 and 999. For this recognition test we used a data base of 26 speakers (14 men and 12 women), each speaker having recorded the 10 digits, 50 random numbers between 00 and 99 and 50 between 000 and 999. Half of this data base was used in the study of the acoustical realizations. This data base, containing about 2900 numbers, was separated into 2 parts. The data from 13 speakers were used for training the model parameters, and the data from the 13 other speakers were used for measuring the recognition performances in a speaker independent mode. The acoustical parameters used are the Mel frequency cepstrum coefficients, plus the total energy and its temporal variation. They are computed every 20 ms (frame rate) using the energy in 24 Mel filters; the bandpass of the signal being 6.4 hKz.

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, independently of the context. However, because of strong coarticulations effects, the sequences /t.r/, /v.a/ and /y.i/ were considered as basic units and thus were represented by a single acoustical model. Using this description, we achieved 93.1 correct recognition rate on the whole numbers for the testing set. Using an average of 2 allophones by phoneme, introducing specific models to handle transitions between adjacent vowels, and keeping the 3 "clusters" mentionned above, we achieved 94.9% correct recognition rate on the same data base, thus reducing the error rate by 25%.

CONCLUSION

This paper shows that a good description of the vocabulary improves 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 various speakers. However, it seems that to correctly predict all the coarticulations, it would be necessary to consider, besides the immediate context, the individualities of the speakers and also the speaking rate of the current sense-group. The set of allophones, defined for this specific application, can easily be extended to fit new vocabularies.

Although the current version of our speech recognition system cannot handle segment duration information, we noticed that the duration is an important cue for differentiating a final devoiced /z/ from /s/. An extra cue for identifying a final devoiced /z/ is the realization of a schwa after the fricative.

BIBLIOGRAPHY

[1] D. Jouvet, J. Monné, D. Dubois: "A new network-based speaker-independent connected-word recognition system"; IEEE proc ICASSP 1986, Tokyo, pp 1109-1112, April 1986.

[2] P. Delattre, A.M. Lieberman, F. C. Cooper: "Acoustics loci and transitional cues for consonants"; JASA, No 27, pp 769-773, 1955.

[3] E. Emerit: "Nouvelle contribution à la théorie des locus - Première partie" Phonetica, Vol 30, No 1, pp 1-31, 1974.

[4] M. Rossi, Y. Nishinuma, G. Mercier: "Indices acoustiques multilocuteurs et indépendance du contexte pour la reconnaissance automatique de la parole"; Speech communication, North-Holland, pp 215-217.

[5] P. Delattre: "Durée vocalique et consonnes subséquentes"; Le Maître Phonétique, 67, 3-ième série, London, 1939.

[6] F. Dell: "L'accentuation dans les phrases en français" in "Les représentations en phonologie" (F. Dell, D. Hirst, J.R. Vergnaud), Paris, Herman, 1984.

[7] A. Di Cristo: "De la microprosodie à l'intonosyntaxe"; Thèse de doctorat d'Etat, Université de Provence, 1978.

[8] K. Bartkova, C. Sorin: "Predictive Model of Segmental Duration in French"; (109-th ASA Meeting), JASA, Suppl 1, Vol 77, p S54, Spring 1985.

[9] F. Lonchamp: "Phonétique et phonologie"; Formation au traitement de la parole, Fascicule 1, Institut National Polytechnique de Grenoble, Grenoble, 1986.

[10] J. Vaissière: "Variance and invariance at the word level" in "Invariance and variability in speech processes", edited by J.S. Perkell and D.H. Klatt, Hillsdale, New-Jersey, London.

[11] R. Jakobson, C. Gunnar, M. Fant, M. Halle: "Preliminaries to speech analysis: The distinctive features and their correlates"; The MIT Press, Cambridge, Massachusetts, 1969.