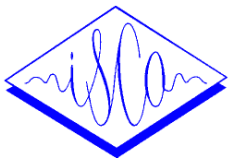


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

PHONETIC TRIPLETS IN KNOWLEDGE BASED APPROACH OF ACOUSTIC-PHONETIC DECODING

Yves LAPRIE Jean-Paul HATON Jean-Marie PIERREL

CRIN/INRIA
BP 329
54506 Vandœuvre-lès-Nancy FRANCE

ABSTRACT

We propose an original knowledge based approach of acoustic-phonetic decoding of continuous speech relying on the use of triplets: *a phone with its phonetic context*. A triplet is made up of two description levels: an acoustic description in terms of acoustic events (formant, burst ...) and an expert component which indicates the significant acoustic correlates to recognize a triplet.

The matching process is firstly performed at the acoustic level, results are then weighted with the help of the expert component. As triplet are phone prototypes it is possible to extract acoustic relations between two reference triplets. These relations must be simultaneously satisfied between two triplet instances of the unknown sentence and two reference triplets proposed as solution. Relaxation techniques enable to implement this idea in order to increase consistency of the global solution.

1 introduction

The past five years have witnessed a large success of Hidden Markov Models in continuous speech recognition. This success mainly originates in the fact that they enable to modelize the speech signal without needing any fine knowledge of speech production phenomena. This absence of knowledge certainly makes possible good global performance rates for an average application (about 1000 words for the SPHINX system [5]) but makes highly difficult further improvements.

Several spectrogram reading experiments ([2] [6]) have proved that an expert identifies more than 80% of phones from a sentence; that is better than the phone level of a HMM based system. On the other hand expert systems in spectrogram reading developed since the early eighties do not equal these performances. CRIN has designed and implemented such a system APHODEX [1] with the help of F. Lonchamp. Knowledge was represented by production rules. This system proved fairly effective to simulate expert reasoning. On the other hand it appeared rather difficult to assess the knowledge amount of APHODEX; this difficulty is increased by the large number of rules which must be added to correctly modelize contextual phenomena of continuous speech.

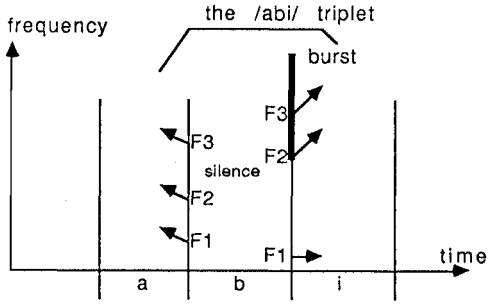


Figure 1: Structure of a triplet

2 Definition of triplet

As acoustic-phonetic knowledge is intended to classify phones it is quite normal to construct prototypes which represent phones to be recognized. In order these prototypes can be usable they must modelize acoustic realization by taking into account coarticulatory phenomena generated by neighboring phones. That led us to propose an approach whose knowledge grain is the triplet:
a phone with its phonetic context

Note that a triplet must not include the whole of neighboring phones or else triplet becomes subject to contextual phenomena we want to escape. Triplet is structured on the two boundaries of the central phone (fig. 1). Rather than an accurate limit which is often not possible to locate, boundaries should be regarded as the area where most of the coarticulatory effects appear. The following acoustic events may be bound to boundaries:

- formant transitions,
- burst described by duration, intensity and main energy concentrations,
- friction noises with low noise limit and intensity. The friction noise of a fricative triplet is divided into two parts; one bound to the left boundary, one to the right one. This artificial splitting enables to represent upward and then downward low noise limits as well as friction noises following bursts.

In order that frequency references are accurate enough we have added a "triplet center" which indicates the formant frequencies in the middle of the triplet.

As we have been defining it, a triplet is only an acoustic description of a contextual phone and thus it does not content any information regarding the knowledge accumulated by an expert in spectrogram reading. The acoustic description of triplet is thus supplemented by acoustic correlates used by a spectrogram reader. An acoustic correlate relates acoustic events fitting the same articulatory event. Furthermore it must not depend on speaker, must be resistant enough and must be acknowledged as significant by the expert, here are some examples:

- "velar pinch" (F2 and F3 drawing closer at the beginning of velar plosives),
- location of energy concentrations of burst in relation to the location of formant trajectories.

The two components of this representation enable to direct phoneme recognition either towards triplet identification from acoustic correlates when they appear clearly enough in speech signal, or towards triplet identification from acoustic realization in the other case. The expert component is very interesting because it helps to recognize with higher certainty triplets which content characteristic acoustic features.

More than a knowledge representation for acoustic-phonetic decoding triplet offers a simple way to adapt phoneme recognition according to the speaker. It is of course possible to perform a frequency normalization on triplets belonging to the knowledge base; this normalization should be fairly coarse since it needs to be fast. Actually one can take advantage of triplet in a more effective way: the idea is to insure that frequency relations between two triplets of the knowledge base which have been proposed as solution for two unknown instances of triplet, are still verified by the two instances. We will come back to this point which relies on the unicity of speaker during the sentence to be decoded in section 4.2.

3 Knowledge acquisition

Amount of knowledge to be learned is obviously considerable even if many triplets are not found in French. Tubach and Boe have shown that 1750 triplets represent 75% of triplets from a very large corpus.

Description of triplet is performed with the help of a user-friendly triplet editor which has been developed on the kernel of the Snorri system [4]. The description is directly achieved on the spectrogram representation (fig. 2); the user draws with mouse the acoustic events belonging to the triplet to be described. The user can take advantage of the automatic acoustic event detectors (automatic formant tracking for example); if need be the user can correct the acoustic description. As the important thing is that the reference triplets are built from correct data we prefer at present to keep this interactive learning which insures the relevance of the built triplets. In order to construct a triplet as representative as possible the user can extract all the occurrences of a given three phone sequence from a corpus; that prevents of retaining erroneous acoustic events. In the second part of learning the knowledge of the expert takes place in retaining the acoustic correlates which are useful in triplet recognition.

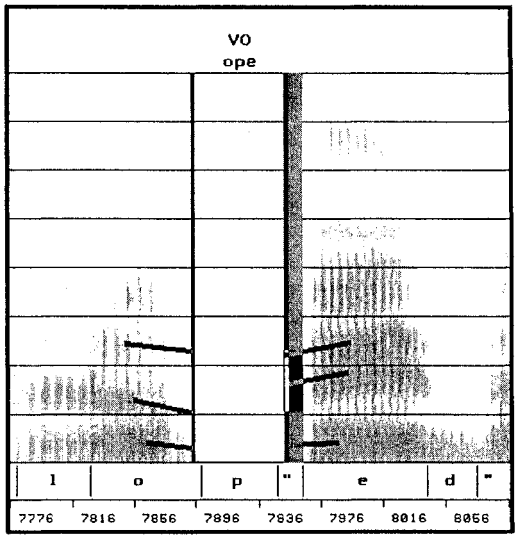


Figure 2: Construction of the /ope/ triplet

4 Triplet identification

Triplet recognition process described in this section comes after segmentation (the segmentation in phonetic classes is achieved by the algorithm designed by D. Fohr [3]) and detection of acoustic events. These two steps are intended to build from the unknown sentence a sequence of triplet instances which will be identified with the help of reference triplets.

Triplet identification is divided in two successive steps. The first one consists in proposing for each instance of triplet the list of reference triplets which match the best the instance. The second step is intended to increase the consistency of the global solution for the sentence. This step insures that constraints existing between triplets of the knowledge base, which have been proposed as solutions, are verified by triplet instances of the sentence to be decoded.

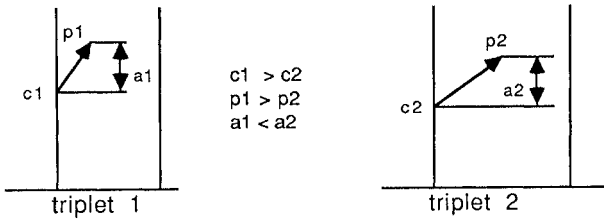
4.1 Coarse labelling

This labelling is a local triplet recognition which relies on matching the unknown triplet with the triplets of the knowledge base. Note that as the phonetic class of instance is known only the reference triplets of the same class are concerned.

One of the key point of prototypes based system is the knowledge organization which should not require more than a limited scanning of the knowledge base. The solution which is commonly adopted in vision is to hierarchize pro-

totypes according to their characteristic features. As we do not know a priori all the possible acoustic correlates which may occur in speech we cannot use such an hierarchy. We have thus chosen to organize triplets according to their formant values.

This organization type influences the matching process which begins by comparing the acoustic description of triplets. The results of this matching are then weighted by taking into account the expert component of triplet. The acoustic matching relies on the computation of a coarse acoustic distance between two triplets. With regard to formant for example, this distance mainly represents the cost of displacement of formant trajectories of the first triplet towards the trajectories of the second triplet.



(Only triplet boudaries and F2 are represented)

Figure 3: Partial constraint on two formant trajectories

4.2 Consistency of the global solution

Results of the preceding coarse labelling is only the juxtaposition of local solutions proposed for each instance of the sentence to be decoded. This does not insure that the global solution is consistent for the whole sentence. It means that reference triplet labels resulting from the preceding labelling may not verify the constraints between two instances. A constraint relates formant trajectories of two triplets (fig. 3). Eliminating candidates for a segment which are not consistent with the most reliable candidates for other segments in the decoded utterance increases the global consistency.

The problem to be solved is as follows:
$N = \{i, j \ldots\}$ is the set of nodes (speech segments)
L_i is the set of labels proposed for the node i
How to eliminate labels which are inconsistent with labels of other nodes ?
This is the classical scheme of discrete relaxation, known under the name of Waltz filtering [9]. There is two types of consistency:

- global consistency: constraints are simultaneously satisfied for each label of every node,
- arc consistency: this consistency is weaker than the preceding one and insures that labels of two mutually constrained nodes are consistent with each other.

We are concerned by the second kind of consistency. Several algorithms have been designed to solve this problem.

Constraints (fig. 3) are built on formant trajectories and take into account three criteria:

- relative location of formant trajectories at triplet boundaries ($c_1 > c_2$),
- relations between slopes of formants transitions ($p_1 > p_2$),
- relation between frequency amplitude of transitions ($a_1 > a_2$)

The fact that some speech segment are either not correctly located or altered by noise may leed to an empty global consistent labelling; it is thus necessary to have at disposal a weak relaxation algorithm. We have chosen AC4 [7] [8] (developed in CRIN for computer vision) and its weak version which prevents from eliminating all the labels.

This relaxation step after coarse labelling has two main advantages:

- it enables to simulate the behavior of a spectrogram reader who proposes a consistent solution for the whole sentence,
- frequency relations imposed by constraints spare an accurate frequency comparison for the coarse labelling.

Fig. 4 shows an example of triplet recognition.

5 Performance evaluation

For the moment we have only tested our approach on a small number of vowels (25 vowels for 10 male speakers belonging to a french corpus). As the number of vowels was very limited we have taken into account only the central phone of triplet for the matching process. Only coarticulatory effects which are strong enough to influence center of triplet are thus considered. Table 1 shows the performance rate. Most identification errors originate in a formant tracking error. The performance is quite satisfactory considering that this system is the first implementation of our approach.

38%	correct triplet in first position
36%	correct triplet in top 3
17%	correct vowel (with incorrect context) in top 3
9%	error

Table 1: Performance rate

6 Conclusion

Many applications of automatic speech recognition system concern limited vocabularies. As the number of triplets for such an application is limited it is thus conceivable to design a triplet based phonetic recognition system which is speaker independent. Learning phase is still an important task but becomes feasible comparing to the one required by recognition of the whole language.

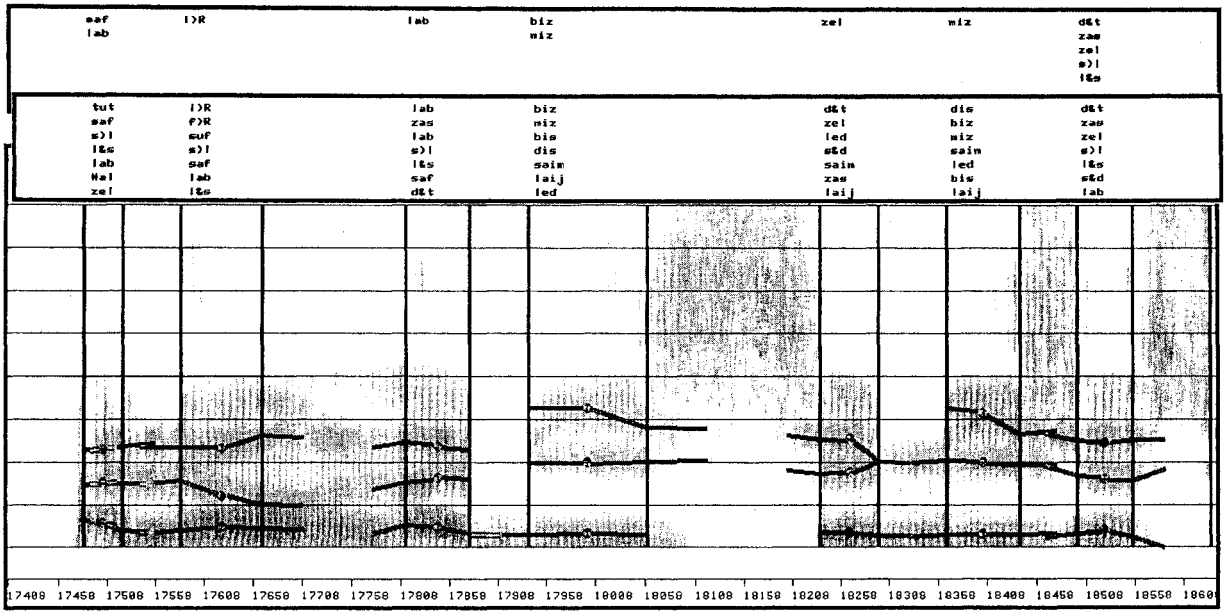


Figure 4: From the bottom up: spectrogram and instances of triplet with formant trajectories, coarse labelling, labelling after relaxation

At present our system uses the segmentation algorithm which has been developed for the APHODEX system. We are thinking of adding a module which can propose several segmentations in case the APHODEX segmentation seems erroneous.

The advantages of triplet as phonetic unit originates in the fact that triplet is a complete information unit. It is thus very easy to manipulate and to deform triplets in order to simulate speaker normalization or to modelize other phenomena like speech emphasis.

References

[1] N. Carbonell, J. P. Haton, D. Fohr, F. Lonchamp, and J. M. Pierrel. APHODEX, design and implementation of an acoustic-phonetic decoding expert system. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986.

[2] R.A. Cole, A.I. Rudnicky, and V.W. Zue. Performance of an expert spectrogram reader. *J. Acoust. Soc. Amer.*, 65, Supp. 1:S81, Paper presented at the 97th meeting of the ASA, 1979.

[3] D. Fohr. APHODEX : Un système expert en décodage acoustico-phonétique de la parole continue. *Thèse de Doct. Univ. de NANCY 1*, 1986.

[4] D. Fohr and Y. Laprie. Snorri: an interactive tool for speech analysis. In *Proceedings of European Conference on Speech Technology*, Paris, France, September, 1989.

[5] K.F. Lee. *Large-Vocabulary Speaker-Independant Continuous Speech Recognition: The SPHINX System, PhD thesis.* CMU-CS-88-148, Carnegie-Mellon University, April 1988.

[6] F. Lonchamp. Reading spectrograms: the view from the expert. In J.P. Haton, editor, *Fundamentals in Computer Understanding: Speech and Vision*, Cambridge University Press, 1987.

[7] R. Mohr and C. Henderson. Arc and path consistency revisited. *Artificial Intelligence*, (28):225–233, 1986.

[8] R. Mohr and G. Masini. Good old discrete relaxation. *Proc. ECAI*, 1988.

[9] D. Waltz. Understanding line drawinggs of scenes with shadows. In *The Psychology of Computer Vision*, chapter 2, pages 19–91, McGraw-Hill, New York, 1975.