

On The Use of A Taxonomy of Time-Frequency  
Morphologies for Automatic  
Speech Recognition

Renato De Mori and Mathew Palakal

Concordia University, Department of Computer Science,  
1455, de Maisonneuve Blvd, Montreal, Quebec. H3G 1M8

Abstract

A computer vision approach based on skeletonization and hierarchical description of speech patterns is proposed. Learning hierarchical descriptions of phonetic events is discussed. Experimental results are reported showing the power of the approach in the recognition of diphthongs in connected letters and digits.

1. Introduction

Most of the popular techniques used today for Automatic Speech Recognition (ASR) are based on comparisons between prototypes and speech data [1]. When the recognition task involves simple vocabularies, word prototypes are used. For complicated tasks, syllabic prototypes or centisecond prototypes are used [2]. In the latter case, a comparison between data and prototypes allows one to assign a label to segments of speech having fixed duration. The label of a speech segment is one of the prototype that best matches that segment. Matching is usually context-free.

Whether knowledge about speech analysis, synthesis and perception should be taken into account or not in ASR is still the object of discussions among the researchers in the field. Automatic recognition of connected spoken letters and of large vocabularies is still an unsolved problem.

As an attempt to solve this problem, a system of plans for extracting and using acoustic properties has been proposed and a general framework for its implementation has been described [3].

The system allows to segment continuous speech into pseudo-syllabic segments. Each segment is not necessarily a syllable, but an acoustic unit to be described. Some portions of this unit can act as contextual constraints for the description of other portions.

The purpose of this paper is that of introducing a novel approach for the description of acoustic segments characterized by spectral lines.

A skeletonization algorithm is applied to digital spectrograms. A variable number of lines with different durations inside an acoustic segment are thus obtained avoiding the errors and the difficulties of tracking formants. A pattern of spectral lines is represented by a hierarchical description.

For the application described in this paper there are only four levels in the hierarchy taxonomy but the levels

as well as the relations at each level can be expanded in order to make the taxonomy reliable enough for a given recognition task.

Experimental results on the characterization of diphthongs in connected digits and letters are discussed.

2. A Taxonomy for Spectral Lines

Spectral lines are extracted with a skeletonization algorithm from the time-frequency-energy patterns obtained by considering the 0-4 kHz portions of spectra computed with the Fast Fourier Transform (FFT) algorithm applied to the preemphasized speech signal. A hierarchical description of spectral lines is then obtained.

The time-frequency-energy pattern for a given pseudo-syllabic segment is processed by a skeletonization algorithm whose details are given in [4].

2.1 The description hierarchy for spectral lines.

The description hierarchy for spectral lines is based on acoustic properties that are known or are expected to be perceptually significant.

The hierarchy follows an open taxonomy that can be expanded to incorporate new items and new classes.

At level-0 of the taxonomy spectral lines are described by vectors  $V_j$  of triplets  $(t_j, f_{ji}, e_{ji})$  ( $j = 1, \dots, J; i = 1, \dots, I$ ) where  $t_j$  is a time reference in centiseconds,  $f_{ji}$  is a frequency value in Hz and  $e_{ji}$  is an energy value in dB.

At level-1 spectral lines are described by morphology symbols  $x_j \in \Sigma_1$  and a sequence of attributes, consisting of time and frequency values.  $\Sigma_1$  is an alphabet obtained by concatenating two symbols belonging to alphabets  $E_1$  and  $\Sigma_1b$ .  $\Sigma_1a$  describes temporal events and is defined as follows:

$\Sigma_1a : \{A:ascendent, H:horizontal, D:descendent\}$

$\Sigma_1b$  gives a rough indication of the frequency location of the mid-point of the line:

$\Sigma_1b : \{LO:low, LA:low-average, A:average, All:average-high, lii:high, VH:very-high\}$

Notice that level-1 descriptions contain pointers that allows one to exactly pick-up the triplets of values at the level-0 description.

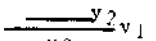
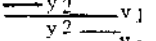
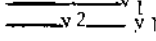
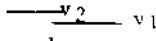
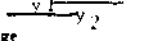

Level-2 descriptions refer to local temporal relations of level-1 descriptions representing lines that are close in frequency. They are of the following type:

$$b_m : R_m (y_{m1}, y_{m2})$$

where  $R_m$  is a relation symbol,  $y_{m1}$  and  $y_{m2}$  are line descriptions like  $a_k$ .

$R_m$  symbols belong to an alphabet  $\Sigma 2$  whose elements are defined in Table I.

Table I

Symbol	Description	Definition
$I(y_1, y_2)$	$y_1$ includes $y_2$	
$LI(y_1, y_2)$	$y_1$ includes on the left $y_2$	
$RI(y_1, y_2)$	$y_1$ includes on the right $y_2$	
$FI(y_1, y_2)$	$y_1$ follows $y_2$	
$FD(y_1, y_2)$	$y_1$ follows down $y_2$	
$FU(y_1, y_2)$	$y_1$ follows up $y_2$	
LCL	clusters in low frequency range	
MCL	clusters in medium frequency range	
HCL	clusters in high frequency range	

Level-3 descriptions capture important frequency relations in a broad frequency range of level-1 and level-2 descriptions. They are of the type:

$$C_n = Q_n(Z_{n1}, Z_{n2})$$

where  $Z_{n1}$  and  $Z_{n2}$  can be level-1 or level-2 descriptions.

$Q_n$  symbols belong to an alphabet  $\Sigma 3$  that contains, at the moment, three symbols.

$\Sigma 3 : \{BF: \text{the back feature, CF: the central feature, and FF: the front feature}\}$

In this way, spectral morphologies relevant for computer perception are extracted and described through subsequent levels of abstraction without losing detail of the original spectra. Fig. 1 shows an example of the description obtained for the diphthong /æi/ in /k/.

### 3. Hypotheticals Generation Using Hierarchical Descriptions

Expressions of predicates whose arguments are elements of hierarchical descriptions are used as preconditions for actions of various types. Some actions consider, for example, the concatenation of spectral lines through a low energy transition with possible gaps as in the case of /ju/ as represented in Fig. 2. Here the detection of an FF followed by a BF is a precondition for searching the above mentioned low energy transition which is circled in Fig. 2. Once descriptions at all levels have

been obtained, then specific parameters relating elements of different descriptions can also be extracted and a-priori probabilities of them can be collected.

The entire descriptor can be seen as an expert system that contains a set of operators. Operators are clustered and clusters are ordered so that there will be a cluster of operators for each level of descriptions. Possible chains of operators are specified by the Expert System Knowledge. The specific chain of operators that is applied on a given pattern depends on the matching between Knowledge and data.

Vector Quantization (VQ) can be considered as one of the operators making this system more general than the ones just based on VQ.

Part of the Expert's Knowledge is used for hypothesis generation and may contain a-priori probabilities.

Let (k) be the hierarchical description of the k-th syllabic segment. Hypothesis H(k) are generated by matching  $\Phi(k)$  with the system knowledge.

H(k) may contain ambiguous hypotheses. For example, a vowel can be identified as a front vowel, but a doubt may remain whether the vowel is /i/ or /e/. Hypotheses are linked with the descriptions that generated them and a summary about hypotheses and descriptions is kept.

As k increases, the summary is updated and consistencies are evaluated and used for pruning ambiguous hypotheses.

For example it is well known that spectral lines are related to formant frequencies and that formant frequencies of vowels are among the acoustic properties mostly affected by speaker variability. In the case of connected pronunciations of letters, it is difficult to distinguish between /i/ and /e/ until a diphthong /aci/ is hypothesized. At this point, the system control knows better what are the differences between /e/ and /i/ for a particular speaker and can put this knowledge into the summary and use it for disambiguating hypotheses already considered. The knowledge written into the summary about a speaker or its mood remains in the summary frame until new evidence makes it change.

The idea of using a summary frame for checking consistencies has been applied to the simple example that will be described in Section 4. It appears to be promising especially because relaxation methods are applicable inside the summary frame thus allowing to check consistencies among acoustic data of the same speaker collected in a short period of time.

Maintaining in time the belief contained into the summary frame is not an easy task because it is difficult to establish when a belief has to be considered obsolete. It is certainly worth keeping into the summary frame acoustic information collected in a frame during the pronunciation of a sentence.

A system for inductive learning of discriminant descriptions involving acoustic properties of phonetic events has been developed [5]. This system is based on principles presented in [6] and can be adapted to the case of hierarchical descriptions.

**4. Experimental Results and Conclusions**

The experiment reported in this section refers to the use of hierarchical descriptions for improving the recognition accuracy of connectedly spoken letters belonging to the so called EI set defined below:

$$EI = \{E, G, P, \beta, V, K, C, B, T, D\}$$

From previous experiments reported in [3], the confusion between /k/ and the other elements of the set is responsible for more than 10% of the overall error rate. As /k/ is the only letter containing a diphthong, it is expected that the detection of the diphthong /æi/ will improve the recognition of the EI set.

For this purpose samples from 5 anglophone speakers were selected (3 male and 2 female); each one pronounced 20 sequences of five letters each.

Knowledge for the /æi/ of /k/ contains an FF with a descendent line or a follows-down feature on the first element of the relation and an ascendent line or a follow-up feature on the second element. For the other letters, the /i/ vowel was characterized by the FF feature without the follows-down or descendent feature considered for /æi/.

The results shown in Table II allows to reduce to zero the errors involving the letter /k/.

**Table II**

		Features				
Speakers	Letter	FD	FF	$\Delta F1/\alpha/$	$\Delta F1/i/$	$\Delta F2$
#1	/k/	100%	100%	425-450	325-345	1950-2250
	others	0	100%	-----	345-525	2125-2375
#2	/k/	100%	100%	475-525	350-375	2125-2250
	others	0	100%	-----	425-525	2175-2350
#3	/k/	100%	100%	275-450	300-325	1800-2200
	others	0	100%	-----	300-500	2100-2350
#4	/k/	100%	100%	575-650	325-350	2025-2300
	others	0	100%	-----	355-510	2100-2325
#5	/k/	100%	100%	475-525	275-300	2175-2400
	others	0	100%	-----	325-550	2200-2475

The strong evidence of follow-down feature and 100% presence of FF in letter /k/ allows to distinguish the diphthong /æi/ and to unambiguously recognize the letter /k/. Table II shows also the frequency intervals in which spectral lines of /æ/ and /i/ involved in the FF relation were detected. As these intervals overlap, it appears doubtful that context-free recognition algorithm can be efficient in a multi-speaker detection of the diphthong /æi/ as opposed to the vowel /i/.

Pronunciations of connected diphthongs in {1, I, U, 0, 5} pronounced connected were analyzed. Sentences from 5 male and 5 female speakers were considered.

Temporal relations involving disjunctions of conjunctions of descriptions were inferred. These relations allowed to correctly segment and unambiguously characterize more than 90% of the data. Work is in progress for disambiguating the most difficult data and for analyzing more speakers.

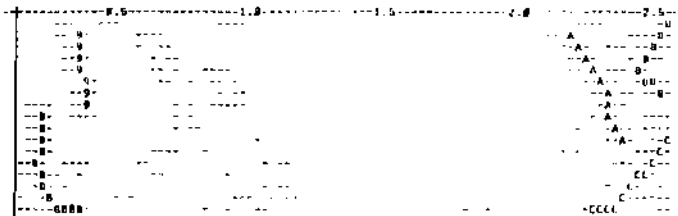
The results obtained show that heirarchical descriptions are powerful tools for detecting and recognizing diphthongs as opposed to single vowels. The research will continue towards the goal of the multi-speaker recognition of connectedly spoken letters and numbers.

Acknowledgements

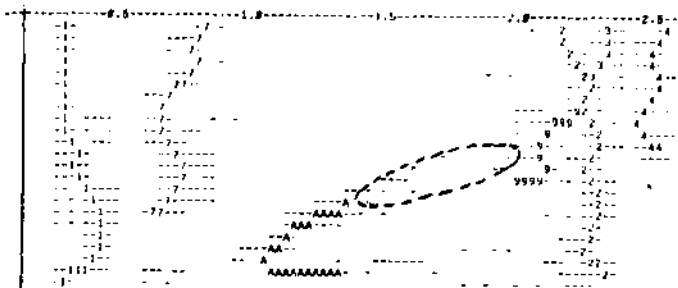
This work was supported by the Natural Science and Engineering Council of Canada.

Bibliography

- [1] L. R. Rablner and S. E. Levinson  
Isolated and Connected Word Recognition. Theory and Selected Applications, IEEE Transactions on Communications, vol. COM-29, no. 6, pp. 521-550, 1981.
- [2] L. R. Bahl, F. Jellnek and R. L. Mercer  
A maximum likelihood Approach to Continuous Speech Recognition, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. PAMI-5, no. 2, pp. 179-190, 1983.
- [3] R. De Mori, P. Laface and Y. Mong  
Parallel Algorithms for Syllabic Recognition in Continuous Speech, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. PAMI-7, no.1, January, 1985.
- [4] R. De Mori and M. Palakal  
On The Use of Computer Vision Techniques for Automatic Speech Recognition, proc. IEEE-CVPR'86, LA, 1986.
- [5] R. De Mori and M. Gilhoux  
Inductive Learning of Phonetic Rules for Automatic Speech Recognition, Proc. CSCSI-84, London, Ont., pp. 103-105, 1984.
- [6] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell  
eds., Machine Learning: An Artificial Intelligence Approach, Tioga Press, Palo Alto, CA, 1983.



**Fig 1. Spectrogram Skeleton of /æi/ in /k/.**



**Fig 2. Example of Concatenation of Spectral lines.**