



ACOUSTIC DISCRIMINATION AMONG WORDS BASED ON DISTANCE MEASURES

P. D'Orta¹

ABSTRACT

In the development of a large-dictionary real-time speech recognition system, an approach commonly accepted is based on a multi-stage design (ref 1). In the first stages, starting from the acoustic data produced by uttering an item (syllable, word, sentence), a fast selection of a small subset of the vocabulary is performed. In the last stage, a detailed search of the most likely item is conducted over the previously identified subset. The selection, as fast as possible, should be able to include always the pronounced item; nevertheless, it must have a high resolution power, that is keep small the chosen subset.

We approach the design of one of the stages by the introduction of classes of equivalence among items, selected via the definition of an acoustical distance. Each item (a word in our case) is represented by a hidden Markov model (HMM), giving a statistical description of the relationship between words and acoustical data. We investigate two different definitions of distance between words: the first one identifies the capability of the model of a word of producing the acoustical data generated by uttering several instances of another word; the second definition is based on differences in the structure and parameters of the models of words.

Starting from the obtained distance matrix, a classification method is used. It is based on a minimal spanning tree approach and allows to find the classification which could keep low the number of words to be selected for the following detailed phase.

INTRODUCTION

The introduction of classes for speech recognition has been already investigated; different solutions have been proposed regarding small-vocabulary speaker-independent systems (ref 2), as well as large-vocabulary systems (ref 3, 4). Classification can be performed automatically or starting from knowledge rules. In a previous work (ref 5) we studied the capability of building word classes from phonetic categories and compared the obtained classifications to knowledge defined ones in a speech recognition system for the Italian language. In this paper a different approach to classification is discussed, based on the study of the properties of hidden Markov models of words.

Each word in the dictionary is represented by a HMM, built from the concatenation of HMMs of the phonemes which compose the word. Parameters of HMM are estimated via a maximum likelihood training procedure based on the Baum-Welch algorithm (ref 6).

The evaluation of the word distance matrix can be thought as speaker dependent. Nevertheless, this could be computationally too heavy for the complete speaker enrollment procedure, so we chose an approximated speaker independent measure to be evaluated in advance. This has been achieved performing the model parameters estimation from speech data collected by several speakers, giving the representation of an *average speaker*. Therefore, the distance matrix is computed only once and is the same for all the speakers.

The first approach followed in the evaluation of the distance matrix starts from the collection of several utterances of all the words. The distance between two words depends on the difference of probability of the models of the two words to produce the collected data. In the next section we will describe in more detail this concept.

A problem arises when the size of the dictionary grows. In fact, being the amount of utterances needed linearly dependent on the dictionary size, it could be difficult to collect the necessary large amount of speech data. Moreover, using this method it is not possible to evaluate the distance between words whose utterances are not available. This implies that it is not immediate to estimate the distance matrix for other dictionaries, even with only a few new words. This problem, although not so relevant, can be solved using the second approach. The distance between two words is evaluated looking at the structure of the models of the words, that is on the basis of the phoneme sequences which form them. The complete distance matrix can thus be computed only from the knowledge of HMM of words. In the third section we will describe this approach.

The evaluation of the word distance matrix is the first step of the classification procedure. In the fourth section we will discuss a clustering technique which, starting from the distance matrix, is able to divide words into classes. We approach a minimal spanning tree method (ref 7), which presents several advantages with respect to other techniques.

Finally, in the last section, we will briefly outline possible applications and future developments of this classification method.

¹ IBM Italy, Rome Scientific Center, Via Giorgione 159, 00147 Roma, Italy

WORD DISTANCE FROM ACOUSTIC DATA

Each word w_k in the dictionary is represented by a HMM, built from the concatenation of HMMs of the phonemes which compose the word:

$$w_k = Ph_{k1}Ph_{k2} \dots Ph_{kn_k} \quad (1)$$

A Markov model of a phoneme Ph^i describes statistically the emission of acoustic data during the pronunciation of sounds associated to the phoneme itself. In our system (ref 8) phonemes are defined by discrete Markov sources which take a transition from a state to another and emit a symbol every 10 msec.. Symbols (*acoustic labels*) belong to a codebook of 200 elements, representing prototypes of like-energy vectors, which are determined during the training session by means of a clustering procedure. In the Markov source, transitions and labels generation are decided on the basis of probabilistic distributions also estimated during the training session.

Given the word w_k , we define $A_k = a_1 a_2 \dots$ the symbol string produced by the utterance of word w_k . If we consider more than one utterance for the word w_k , we will have associate to it the strings $A_{k1}, A_{k2}, \dots, A_{kn_k}$. We define the quantity s_{ki} as the log probability of model of word w_k to produce string A_{ki} :

$$s_{ki} = \log P(A_{ki} | w_k) \quad (2)$$

Given the complete set of strings associated to the word w_k , we define the quantity S_k (*score*) as the average of s_{ki} :

$$S_k = \frac{1}{n_k} \sum_{i=1}^{n_k} s_{ki} \quad (3)$$

Let us now consider a different word w_j . We define the quantity s_{ki}^j as the log probability of model of word w_j will produce the string A_{ki} , that is we measure the capability of the model of word w_j to generate the strings related to utterances of w_k :

$$s_{ki}^j = \log P(A_{ki} | w_j) \quad (4)$$

Similarly, let the score S_k^j the average of the quantities s_{ki}^j :

$$S_k^j = \frac{1}{n_k} \sum_{i=1}^{n_k} s_{ki}^j \quad (5)$$

The distance between words w_k and w_j can then be introduced as:

$$d(w_k, w_j) = |S_k - S_k^j| \quad (6)$$

The function $\delta(n)$ which gives the distance of the n-th element in the list of the closest words from a given one, displays a logarithmic behavior. There is a good resolution power in the discrimination of near words. On the other hand, distant words tend to get high, but very thickened, values. This aspect is not crucial because for the goal of word classification it is important to consider only the closest words.

Computationally, the evaluation of the complete distance matrix becomes heavy when the size of the vocabulary grows. If N is the size of the dictionary, and m the number of utterances per word, it is necessary to evaluate mN^2 terms $P(A|w)$. Let L the average length of each utterance, s the number of states in HMM of phonemes and K the average number of phonemes per word, the computational cost is:

$$C_1 = \alpha_1 m N^2 s K L \quad (7)$$

For a 1000-word dictionary, one utterance per word, the evaluation of the complete matrix took about 20 hours on a general purpose mainframe.

WORD DISTANCE BASED ON MODELS

The computational cost of the preceding approach suggested the investigation of a different definition of word distance. In addition, it could be interesting to have the possibility to evaluate it without the need of a huge set of speech data.

Models of words, characterized by their phoneme sequences, give the intuitive idea that similar structures should reflect low distances and, on the contrary, very different structures should lead to high distances. To extend formally this idea, we

suppose that to each word is associated a new Markov source which is able to emit phonemes. This Markov source results from the concatenation of other Markov sources that we will call *phonemic units*. Each phonemic unit is able to produce phonemes of the language. Given the word w_k , which had associated the model

$$w_k = Ph_{k1}Ph_{k2} \dots Ph_{kn_k} \quad (8)$$

we now consider the Markov source:

$$w_k^P = P_{k1}P_{k2} \dots P_{kn_k} \quad (9)$$

where P_{ij} is the phonemic unit corresponding to the phoneme Ph_{ij} . We define the quantity Z_k^j as the probability of model w_k^P to produce the string of phonemes representing the phonetic structure of model of word w_j :

$$Z_k^j = \log P(w_j | w_k^P) = \log P(Ph_{j1}Ph_{j2} \dots Ph_{jn_j} | P_{k1}P_{k2} \dots P_{kn_k}) \quad (10)$$

Then we define the distance between words w_k and w_j as:

$$d(w_k, w_j) = |Z_k^k - Z_k^j| \quad (11)$$

The evaluation of Z_k^j requires the knowledge of parameters of phonemic Markov sources, that is probabilities of transition between states and probability distributions of emission of phonemes. If phonemes were directly observable from speech it would be easy to estimate these statistics with the usual training techniques. What we need is a way to convert speech into sequences of phonemes. Let $A = a_1 a_2 \dots a_n$ a sequence of acoustic labels representing the utterance of a given text T , that can be described in terms of Markov sources as:

$$T = Ph_1 Ph_2 \dots Ph_r \quad (12)$$

Given the probability distributions of Markov sources Ph^i , $P(a_m | Ph^i)$ that were determined as explained in the previous section, it is possible to align segments of speech to each Markov source in the text T . Let A_l the segment of speech aligned to the l -th phonetic source in the text. To find the phoneme corresponding to the segment A_l , we search for the phonetic source which maximizes the probability of producing A_l :

$$P(A_l | Ph_l) = \max_i P(A_l | Ph^i) \quad (13)$$

To the speech sample A then there will correspond a sequence of *decoded* phonemes:

$$U = Ph_1 Ph_2 \dots Ph_s \quad (14)$$

Considering now the sequence of phonemic Markov sources corresponding to the training text T , and the string of output symbols U , we can train the distributions $P(Ph^j | P^k)$ using the Baum-Welch algorithm.

From a computational point of view, the evaluation of the distance matrix can be represented by the cost:

$$C_2 = \alpha_2 N^2 z K^2 \quad (15)$$

where z is the number of states of phonemic sources, K is the average number of phonemes per word. The ratio of computational costs between the two methods is:

$$\frac{C_1}{C_2} = \frac{\alpha_1 m s L}{\alpha_2 z K} \quad (16)$$

For some experiments, $C_2 = 0.07 C_1$.

This method allows to evaluate distance without the need of speech data, only from the knowledge of the phonetic structure of each word. This is particularly interesting when considering new words, whose pronunciations could not be available.

CLASSIFICATION

Aim of the classification is to divide the dictionary in groups of acoustically similar words. Given a certain dictionary, it is not known the optimal number of classes to be used. Some experiments with K-means clustering techniques showed the inadequacy of an a-priori determination of the number of classes. A method able to dynamically find the number and the composition of clusters is clearly to be preferred.

A minimal spanning tree (MST) approach seems to be adequate to this task. In fact it is able to describe the space of words in terms of strong connections, putting into evidence areas in which words are very close or areas of sparseness. Once built the MST, this must be divided cutting some edges. Edges to be broken should correspond to boundary zones among words, that is where the density of words changes dramatically. This allows to keep together groups of similar words.

The subdivision algorithm is relatively simple. To each edge, representing the connection between two words, their distance (weight) is associated. Then, for each edge, its neighbors are analyzed (usually far not more than two or three steps) evaluating the mean and the variance of their weights.

Let e_k the k -th edge in the MST, we define $w(e_k)$, $\mu(e_k)$, $\sigma^2(e_k)$ respectively as its weight, the mean of weights of close edges, the variance of weights of close edges; the decision rule allows to break edge e_k when

$$w(e_k) > \mu(e_k) + \alpha\sigma(e_k) \quad (17)$$

or when

$$w(e_k) > \beta\mu(e_k) \quad (18)$$

α and β , as well as the number of neighbors of an edge to be considered, are parameters that can be conveniently chosen. These parameters can influence the composition, the size and the total number of clusters.

CONCLUSIONS

The use of word classes is a way to allow fast access to the dictionary when performing speech recognition. Clearly, it must guarantee a good accuracy in the selection of the correct word. Word distance seems to be an easy concept and a valuable mean to make a good division of the dictionary into groups of similar words. Nevertheless, other problems should be solved. A crucial point is, once performed clustering, the choice of the centroid of each cluster. Using one element of the class does not seem to be a good solution; a preferable one will address the identification of an *averaged model* of the words in the class. Besides, some preliminary experiments show that it could be more convenient to make clusters not completely separated. This can avoid some problems due to a not good selection of parameters in the classification phase. Finally, word distance techniques will be investigated in the design of a pyramidal access to the dictionary. Most of these arguments will be object of future communications.

REFERENCES

- [1] T. Kaneko, N.R. Dixon, **A Hierarchical Decision Approach to Large-Vocabulary Discrete Utterance Recognition**, *IEEE Trans. on ASSP*, vol. ASSP-31, no.5, 1983
- [2] S.E. Levinson, L.R. Rabiner, A.E. Rosemberg, J.P. Wilpon, **Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition**, *IEEE Trans. on ASSP*, vol. ASSP-27, no.2, 1979
- [3] D.P. Huttenlocher, V.W. Zue, **A Model of Lexical Access from Partial Phonetic Information**, *ICASSP 1986*, Tokyo
- [4] B. Merialdo, A.M. Derouault, S. Soudoplatoff, **Phoneme Classification Using Markov Models**, *ICASSP 1986*, Tokyo
- [5] P. D'Orta, M. Ferretti, S. Scarci, **Phoneme Classification for Real Time Speech Recognition of Italian**, *ICASSP 1987*, Dallas
- [6] L.R. Bahl, F. Jelinek, R.L. Mercer, **A Maximum Likelihood Approach to Continuous Speech Recognition**, *IEEE Trans. on PAMI*, vol. PAMI-5, no.2, 1983, pp.179-190
- [7] C.T. Zahn, **Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters**, *IEEE Trans. on Computers*, vol. C-20, no.1, 1971
- [8] P. D'Orta, M. Ferretti, A. Martelli, S. Melecrinis, S. Scarci, G. Volpi, **A Speech Recognition System for the Italian Language**, *ICASSP 1987*, Dallas