

SPEECH RECOGNITION BASED ON SPEECH UNITS.

G. ZANELLATO

Faculté Polytechnique de Mons, Bvd DOLEZ, 31 - B-7000 MONS (BELGIUM)

SUMMARY

In a classical quantization system, each vector is represented by the nearest centroid. But it is impossible then to distinguish two vectors belonging to the same class. In order to mitigate this disadvantage, we have taken into account the two nearest neighbours and also a "belonging degree" calculated from the distances between the vector and the two centroids. In the case of speaker independent speech recognition system, this "fuzzy" quantization gives better results.

For the recognition of large vocabularies, it is usual to perform a phonemic segmentation and then to achieve the matching on the obtained sequence. But it is very hard to set up a good modeling for the phonemes. Nevertheless, we can take into account elementary components of the phonemes, for which simplest models may be used. We investigate a set of 100 "acoustic units" that have been defined from the centroids of the fuzzy quantization. The results we obtained are very attractive.

1. INTRODUCTION.

The experience of our laboratory mainly refers to isolated word speaker dependent and speaker independent recognition, in the case of small vocabularies. Nevertheless, we propose a new method allowing to approach the recognition of isolated words (large vocabularies) in a speaker independent way. This principle may possibly be used for the recognition of small sentences or even for continuous speech recognition.

The experimental conditions are the following :

- . anti-aliasing filter cut-off frequency = 4625 Hz. ;
- . sampling rate = 10000 Hz. ;
- . pre-emphasis : $\alpha = 0.95$;
- . frame size = 300 samples + Hamming window ;
- . shift between frames = 100 samples ;
- . number of poles of the autocorrelation analysis : $p = 12$;

The available data base is composed of 100 versions (from 100 different speakers) of 20 isolated french words (10 digits and 10 control words).

2. FUZZY QUANTIZATION PRINCIPLE.

In order to hold a better image of the position of a spectral vector (SV) in the acoustic space, we have developed a new method for the vector quantization :

in the "classical" quantization, each frame is represented by the class number of which the center of gravity (CG) is the nearest. But it is possible to join to this class number a so called "belonging degree" that will characterize the remoteness between the SV and the CG ; moreover, we will take into account the two nearest classes.

So, each SV X_j is quantized by 4 numbers : the two nearest neighbours (C_1 and C_2) and the respective belonging degree (or probabilities) :

$$a_1 = \frac{d(X_j, C_2)}{d(X_j, C_1) + d(X_j, C_2)} \quad \text{and} \quad a_2 = 1 - a_1$$

where

- C_1 is the nearest CG of X_j ;
- C_2 is the second nearest CG of X_j ;
- $d(.,.)$ is the ITAKURA distance [1] between two SV .

One of the criteria used to define the quality of a quantization is the measure of the total weighed distortion (D_T) ; its expression is given by (where N_i is the number of elements belonging to class i and N is the number of classes) :

$$D_T = \frac{1}{N} \sum_{i=1}^N \left\{ \frac{1}{N_i} \sum_{X_j \in (i)} d(X_j, C_i) * a^{(i)}(i) \right\}$$

We shared a spectral space of 52.580 SV (50 versions of the vocabulary) in two ways. On the first hand, with the binary-splitting algorithm (128 classes : QB7) [2], on the second hand, with the proposed method (100 classes : QS100). In the first case, the value of D_T is 0.52 and 0.43 in the second one.

This work has been partly supported by the Belgian Ministry of Economic Affairs under IRSIA-IWONL grant N° 4560.

10.21437/ECST.1987-32

The influence of this method has been tested also at the recognition level. In the case of a speaker-dependent system, there is no change in the results (≈99%). In the case of a speaker independent system, based on the modelization by probabilistic automaton, we used two kinds of models : the "Classical Markov Models" [5] (case 1), and those for which there is no transition matrix (we define two sub-state for each state of the model) but the "showing up" probability concept [6] (case 2). In each case, 50 versions (T1) have been used for the training of the models and 50 different versions (T2) uttered by 50 new speakers, only for the tests. The results are shown in the following table (number of recognition errors for 1000 tests) :

	QB7, case 1	QS100, case 1	QB7, case 2	QS100, case 2
T1	9	11	6	6
T2	34	20	37	17

Those results show directly the interest of the proposed quantization method for the speaker independent applications.

3. RECOGNITION USING SPEECH UNITS.

3.1. INTRODUCTION.

When the size of the vocabulary increases, it is generally agreed to use a phonemic approach for the speech recognition. Unfortunately, this well-known method does not give good results; this mainly is due to the fact that the acoustic sharp of a phoneme is essentially warping and widely influenced by the phonemic context in which it takes place.

It is yet possible to imagine a phoneme as being composed of a sequence of "elementary sounds" (ES), each of which being produced by a sequence of different articulatory configurations of the vocal track. But, we can consider that it exist a one-to-one correspondence between a given position of the vocal track and the position of a SV in the spectral space. Then, the sequence of SV representing an ES draws a path in the spectral space; this path may run across several classes. Different paths may correspond to the same ES. Our objective is to modelize those paths; their set will be called a "speech unit" (SU); this one may then be seen as the image of an ES. Moreover, sequences of different ES may correspond to the same phoneme used in different context.

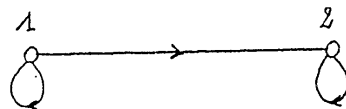
It follows that the choice of the elementary components of a phoneme as base units for the segmentation and the recognition represents an attractive solution. It is also interesting to be able to determine automatically those SU. To this end, we use, as base for the SU, the 100 centroids obtained by the quantization of the spectral space described above.

In brief, this principle allows to segment a vocal signal into a sequence of SU. Those units share the spectral space into "areas" characterized by centroids obtained by vector quantization. Actually, each area may overlap others; they agree with a larger concept than the one of quantization class.

3.2. MODELIZATION OF THE SPEECH UNITS.

3.2.1. INITIALIZATION.

The probabilistic model we retained for a SU is shown in the following figure. The first state characterizes in the best way the sound corresponding to the given SU, i.e. it can be seen as an image of the area of the spectral space corresponding to this SU, and the second one characterizes a sound as "far" as possible from the one represented by the first state; it is, in fact, the complementary area, with regard to the spectral space, of the one associated with the first state.



The thing is to determine the production matrices (there is no transition one). The production probability $b_k(u_k)$ of a class k by the first state of a model is determined in the following way :

- 1°) We use all the quantized SV for which the C_1 corresponds to the SU we want to modelize (for instance, the SU J).
- 2°) Each one of those vectors is regarded to be a part of the class J for an amount equal to its belonging degree a_1 to this class.
- 3°) The sum of those a_1 represents the "amount" (real value) of SV allocated to class J .
- 4°) We also take into account the C_2 (class L) and a_2 of each one of the SV defined in 1°).
- 5°) For each class L , we sum the respective a_2 , what furnishes the "amount" of SV allocated to this class, for the model of SU J .
- 6°) The sum of all those "amounts" (for all the used classes) is equal to the number N of SV defined in 1°).

- 7°) The ratio between the "amount" Q_k attached to the class k and the number N gives the "production probability" of the class k in the first state of the model of the regarded SU.
- 8°) The value of the probability attached to the classes never "used" is half of the lowest probability determined in point 7°).
- 9°) We weight all those probabilities, in order to assign a given value (e.g. 0.9) to the production probability of the class corresponding to the SU we had modeled (i.e. the class J for the SU J). In this case, the sum of the probabilities related to a given state is unitary anymore.

The production probability vector $b_2(u_k)$ associated to the second state of the model is calculated from the first one :

$$b_2(u_k) = 1 - b_1(u_k) \quad k = 1, 2, \dots, 100 .$$

We have then determined the initial production matrix.

3.2.2. TRAINING.

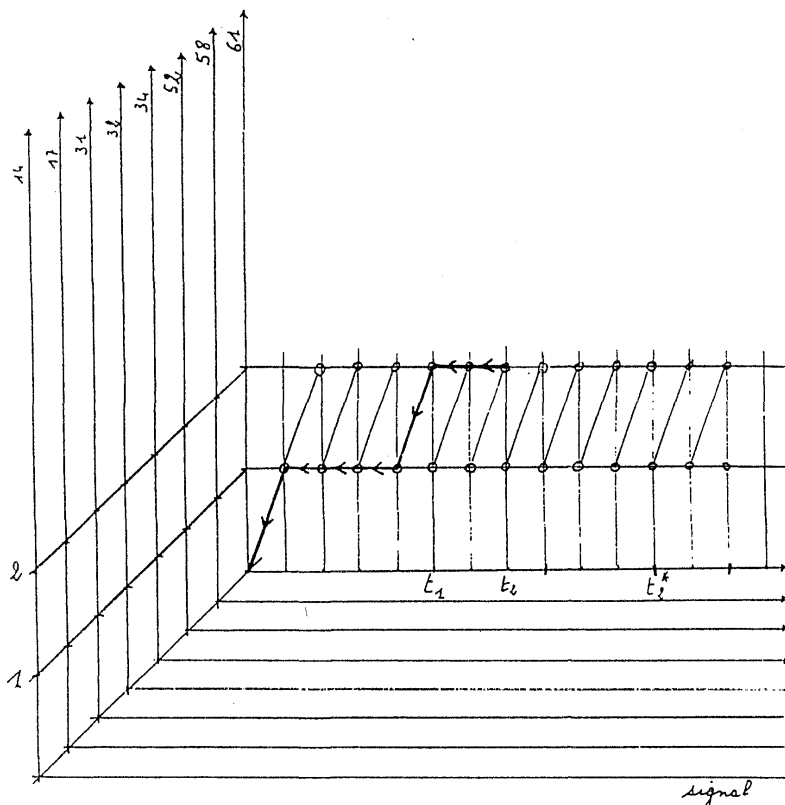
3.2.2.1. RECOGNITION OF CONNECTED WORDS.

The signal we want to recognize is composed of a sequence of "words" (or SU) in any order. It is then necessary to match it with all the reference models, and that must be done from the beginning to the end of the signal. But, the use of the 100 centroids as base for the SU presents an important advantage : the matching will be performed only with the models of the SU really presents in the signal ! (moreover, we process only about 15 frames at a time). For instance, if the signal has the following form (for the nearest neighbour only) :

17-34-34-31-34-34-34-34-52-17-58-58-61-14-32-14-14,

then, the models to be used for the recognition will be those of the SU 14, 17, 31, 32, 34, 52, 58 et 61, i.e. eight models instead of 100 !

The first purpose is to determine the models of SU to be used, then we have to execute several times a kind of Viterbi algorithm (there is one for each retained SU (see following figure)). In each "VITERBI", we have to determine the maximum probability with which we reach the second state. That means we have to execute a "back-tracking" at each frame of the signal (for each model). We consider that we are living the model when we are "looping" for at least three times in the second state.



In the above figure, the backtracking performed from the frame t_1 is such that we may consider still being in the model, but at the frame t_2 , we may consider we are living it (we have chosen 3 running points).

We note the value of the maximum probability obtained in (t_2-3) , i.e. at the moment we "live" the model. We perform the same determination of optimal path for all the models of SU (the t_2 will be different) and we choose as "recognized" the one that in its own (t_2-3) gives the highest maximum probability. Let's suppose t_2^* being this frame. We cut out the frames 1 until (t_2^*-3) , i.e. those that, in the determination of the optimal path, have been assigned to the first state of the "recognized" model. Then we start again the same processing on the fifteen frames following (t_2^*-3) , in order to determine the next SU. Finally, we obtain, for the whole signal, a sequence of SU (let's suppose 34 - 58 - 14 in our example) that represent the segmentation of the word into SU.

3.2.2.2. PARAMETERS CALCULATION.

The SU models will be trained by a set of 1000 isolated words segmented into sequences of SU.

Marking the initial SV groups associated to each SU of each sequence, allows an easy determination of the production probabilities of the first state of each model. Nevertheless, we have to take into account that the contribution of each SV is equal to its "belonging probability" to the regarded class. For instance, if a SV assigned by the training to the first state of the SU model n° 30, is quantized in the following way :

C1 : 17 ; a1 : 0.8 et C2 : 21 ; a2 : 0.2 ,

then, the contribution of this vector is such that the "amount" of the class 17 is increased of 0.8 and the one of class 21 is increased of 0.2.

We take care that any probability has a zero value and then we weight them as explained above.

The probabilities of the second state are calculated from those ones.

3.3. FROM THE SU TO THE PHONEMES.

The second part of the proposed method is related to the identification of the phonemes from the obtained sequences of SU.

We can consider that a phoneme is composed of a sequence of SU drawing a path in the spectral space. The coarticulation effect drives this path across several SU, in regard with the different pronunciations. Taking into account all the possible sequences allows to create a phoneme model. This one will be trained by a set of SU, as well as the SU models were trained by a set of SV.

In order to determine the sequence (or sequences) we use again the words of our data base. Each word of the vocabulary is modeled by a probabilistic automaton composed of a number of states equal to the number of phonemes of the word. The training of those automatons will be performed using the words that have been segmented into SU, in order to determine, for each state, a SU production probability vector. We forecast that after this training, each state of a model will be specialized in the production of a phoneme (i.e. a sequence of SU).

In brief, the proposed methodology consists to forecast a middle stage before the segmentation of a word into phonemes. The main interest is a total automatization of the processing.

The recognition of a word is now performed in the following way : the signal is at once segmented into speech units, performing a mating with the predefined SU models; then the sequence of obtained segments is processed by the automaton representing each of the vocabulary word. The recognition decision is taken at that point.

3.4. RESULTS.

A set of 1000 words (50 versions) has been used for the training of the SU models and of the words models; the error rate is then :

- 1 % (10 errors) if we use the sam words as tests
- 2 % (19 errors) if we use 1000 new words (50 new speakers).

4. BIBLIOGRAPHY.

- [1] F.ITAKURA, "Minimum Prediction Principle Applied to Speech Recognition", IEEE Trans., ASSP-23, February 1975, pp. 67-72.
- [2] R.M.GRAY, "Vector Quantization", IEEE ASSP Mag., April 1984, pp. 4-29.
- [3] H. SAKOE, S. Chiba, "Dynamic Programming Algorithm for Spoken Word Recognition", IEEE Trans., ASSP-26, February 1978, pp. 43-49.
- [4] F.JELINEK, "Continuous Speech Recognition by Statistical Methods", Proc. IEEE, Vol. 64 (April 1976), pp.532-556.
- [5] L.R.RABINER, S.E.Levinson, and M.M.Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent Isolated Word Recognition", BSTJ, V62, N°4, April 1983, pp.10-1105.
- [6] R.BOITE, H.Leich, G.Zanellato "Isolated Word Recognition by Hidden Markov Model", EUSIPCO-86, pp. 541-544.
- L.E.BAUM, T.Petrie, G.Soules, and N.Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", ANN. Math. Stat., 41 (1970), pp.164-171.
- G.ZANELLATO "Quantification vectorielle souple, base de la reconnaissance de la parole", 11 Colloque GRETSI, to be published.