Using Hidden Markov Models in segmentation of speaker-independent connected-digits corpus

FRANCISCO DIAZ, MANUEL RUBIO, PEDRO GOMEZ,
VICTOR NIETO, and VICTORIA RODELLAR
Laboratorio de Comunicación Oral "R. W. Newcomb", Facultad de Informática
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n 28660 Boadilla del Monte, Madrid
SPAIN

Abstract: - The first task to be accomplished in speech recognition is the segmentation and labeling of records. Regarding speech, this is a very complicated and costly procedure, although of most importance because at the present time many available speech corpora are not segmented. This paper proposes a semi-automatic segmentation method in order to reduce the manual segmentation burden of a very large corpus. First, Hidden Markov Models are created with a reduced set of records. Afterwards they are used to perform an automatic segmentation on the rest. Recursively, new more robust models are created and used to create new segmentations. The method consists in three main steps: (1) Initial Reduced Segmentation, (2) Recursive-Extended Segmentation and (3) Post-processing of the labels. This method was evaluated in the segmentation of the TIDIGITS corpus with two independent initial manual segmentations. Finally the method was able to label correctly 96.18 % and 95.72 % of the corpus records, respectively.

Key-Words: - Automatic annotation, Speech segmentation, Markov Models, Speech recognition

1 Introduction

Research in speech recognition, speaker and language identification require the use of corpora whose records contain the variability of the universe of the speakers. Some of the main factors of this variability are gender, age, dialect, recording scenario and others. This implies a high volume of data since all these factors must be well represented. On the other hand, speech recognition experiments cannot be carried out unless information about segmentation and labeling is available. At the present time several available corpora lack such segmentation information, like the TIDIGITS corpus. Speech traces are frequently segmented manually, sometimes with the aid of special tools that simplify the process [1], [2], [3], [4]. There is a wide variety of models for manual labeling accordingly to type and purpose, some are specific for certain databases and others are suited for a specific linguistic domains or applications. Manual labels are very dependent on individual criteria, even when the personnel who create them is trained with identical procedures in spectrogram reading or interpretation and other forms of visualizing and analyzing the speech trace. Besides, manual labeling is a very monotonous and cyclical task that produces a high level of exhaustion in a short period of time and reduces rapidly the precision of the segmentation. Several authors propose automatic segmentation methods considering variations on the measurements of the speech trace computed using short-term algorithms, for example, energy, log-energy, pitch trace, fundamental frequency, and others [1], [2]. These methods are not fully reliable for speaker-independent databases. This fact is due to the individual pronunciation characteristics in continuous speech as the regions of energy variation that may be totally different. For example, intervals of low energy values do not necessarily represent a silence segment between two words, and on the other hand, two connected words without a pause in between do not necessarily generate a region of low energy. It is necessary to consider the characteristics of each class or label in order to perform a correct segmentation.

This paper proposes a method for large corpus segmentation, speaker-independent connected digits, reducing the task of manual segmentation. At the same time, it is intended to reduce the effects of individual criteria while minimizing the costs associated to labeling. The method includes both manual and automatic segmentation. For the automatic segmentation Hidden Markov Models (HMMs) are used in different phases of the process. These are trained in each phase with the correct speech records segmented in the last phase.

The paper is organized as follows: in section 2 the theoretical aspects and possibilities of HMM labeling are analyzed, section 3 describes the proposed method and presents the block diagram of the algorithm, section 4 describes the evaluation of the

method in labeling the TIDIGTS corpus at a word level, finally in section 5 conclusions are derived.

2 Hidden Markov Model-based segmentation and labeling

Research in speech recognition, speaker and language identification require the use of corpora whose records contain the variability of the universe of the
It is well known that left-right Hidden Markov Models may be used in speech recognition tasks [5]. The use of HMMs in speech recognition requires the existence of correctly segmented records, which are used for the training of the different models with the Forward-Backward algorithm or Baum-Welch re-estimation. HMMs do not process directly the speech samples one at a time, but in sequences, after a short-term feature extraction process has been used. This transformation allows using HMMs with continuous probability functions. If the implementation of the models is carried out using discrete probability functions a vector quantization process has to be included. Each HMM model reflects the temporal variations of the templates (in this case digits) to be recognized by means of statistical modeling of the observation states and transitions. Once estimated, the models are used for parsing. The speech records to be recognized are parsed and the HMMs compute the probability of sequence generation using the Viterbi algorithm. Later, the recognized template is chosen according to the model that produced the largest probability. During the parsing phase it is ensured that the winner model has transited through all the model states. For connected digit recognition, the HMMs have to include two more states, one at the beginning and one at the end, in order to connect the models for the analysis of a sequence of templates. The calculation of probabilities in each HMM begin just after a final state has been reached.
The hypothesis justifying the use of HMMs in labeling and segmentation is based in the fact that the instant in which a final HMM state is reached is a possible candidate to be a "contour point" (end of a recognized token and beginning of a new token to be recognized). This scheme is repeated along the digit sequence. Once the parsing has analyzed the entire digit sequence, the contour points determine the segmentation and labels of the segments corresponding to the sequence of winner tokens. This argument is valid for transitions between the templates in intermediate positions of the sequence, but not for the beginning and end of the digit sequence. These points are easily detected applying an end-point detection algorithm, but an extra HMM can also be defined for silence segments, since it represents a class that can be also segmented.

3 Segmentation Scheme

From the analysis presented in the last section it can be asserted that a corpus can be automatically segmented with the use of HMMs if these are trained with enough data in order to correctly model the states and its transitions. However, initially an initial segmentation necessary to assign correct training data to each HMM will not be available. The solution to this problem is to create HMMs recursively, i.e., to carry out an automatic segmentation process and create new HMMs using the last correct segmentation available. This process is repeated until newly produced HMMs do not improve segmentation results anymore.
There is no available correct segmentation reference. Therefore it will not be possible to quantify the quality of the segmentation, at least directly. In this situation, the best alternative is to consider the recognition percentage (rate of correct labeled records) carried out by the parsing phase as a measure of the effectiveness of the segmentation. This conclusion is reasonable since if the recognition is correct, we can assume that the contour points obtained from the token sequence have been well established. However, nothing may actually ensure this, as theoretically it is possible that a well recognized trace might have been incorrectly segmented. Here when recognition is referred the sequence of templates is referred exclusively, not its contour points. Nevertheless, as far as the new HMMs are more general and representative of the template sequence, the precision of the segmentation increases.
The proposed scheme tries to solve all of above-mentioned problem within three basic steps: (1) Initial Reduced Segmentation, (2) Recursive Extended Segmentation and (3) Label Post-processing.

3.1 Initial-Reduced Segmentation

The objective of this phase is to obtain an initial segmentation of the corpus, which may be later improved. The term "reduced" is used since the initial HMMs are created with a very small set of the original corpus.
The phase begins with a random selection of the records with which the first models are to be trained. Later, these records are manually segmented. The number of selected records must not be too large,

although it is desirable that they reflect the variability of the corpus, giving a representation of the set of speakers including several repetitions of all possible labels. After this stage it is necessary to define the topology of the HMMs and to carry out the training. The considerations regarding this task are detailed in the next section. Once the models are produced, they will be validated using the same records with which they were trained, and at the same time a new automatic segmentation is generated based on the contour points generated by the HMMs for such subset of selected records. This new segmentation will substitute the manual one created at the beginning of the process. This substitution allows on one hand to correct possible defects on the manual segmentation, and on the other, disregarding incorrectly segmented records. Thence a new training process is required using the correctly segmented records as evaluated from the last validation phase. The new estimated HMMs are more consistent and will be used to segment the entire corpus. As a result of this step a new segmentation database (information about the contour points and labels for the corpus records) is obtained, which will be called "reduced segmentation". At this point, the segmentation may not be still very accurate and the erroneous segmented data may correspond to a high percentage of the corpus.

3.2 Recursive-Extended Segmentation

This step is carried out in order to enhance reduced segmentation. In this case instead of carrying out a simple segmentation a recursive process is started where new models are estimated using as many templates or records as possible (extended corpus).

Figure 1 shows the block diagram of the recursive-extended segmentation process. In this diagram the blocks with thick lines represent processes and the ones with thin lines represent databases. The process begins taking the vector representation of the speech traces and the segmentation previously calculated ("source segmentation"). The process then determines the correctly segmented data, which is used to estimate new HMMs for those possible labels. Notice that in the first recursion step the source segmentation is the reduced segmentation, which is the result of the previous step. Later the HMMs are validated against the entire corpus and simultaneously a new segmentation (target segmentation) is created. It must be better or similar to the source segmentation since the models from which it is obtained contain a larger representation of the data. The next process is the recognition evaluation, considering that the sequences of labels assigned to the traces correspond exactly with the sequences of HMM winning models. Although we do not dispose of a reference segmentation, the recognition can be evaluated since the sequence of tokens included in each record is known.

For example in the TIDIGITS corpus this information is part of the filenames. Accordingly to the recognition percentage obtained the process can continue with another recursion step or halt. If the number of correct labels is greater than in the previous step the process is repeated, but before, source segmentation is substituted by target segmentation. On the other hand, if there is no increase in the recognition rates the last source segmentation database is selected and the recursive process stops.

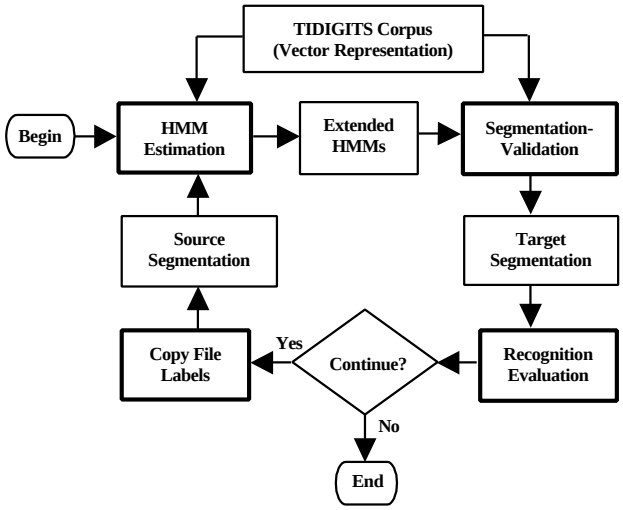

Fig. 1. Block Diagram of the Recursive-Extended Segmentation.

3.3 Post-processing

The last step was designed in order to correctly segment the most of the records. In practice there will always be some data wrongly segmented. Even the correct segmentations might present certain defects, which this last step tries to solve. For example, two or more silence segments can appear consecutively and can be grouped into only one. Regarding incorrectly labeled traces two alternatives can be used to solve them. The first and simpler option is to label them manually, like in the initial-reduced segmentation step. This is preferred if the number of incorrect traces is not too high. If on the contrary the number is elevated, a second option is preferred. It consists in

repeating the method described in this paper for the incorrect subset of traces.

4 Experiments and Results

The proposed method has been used in the segmentation of the Speaker-Independent Connected-Digit Corpus TIDIGITS. This database is used for the design and evaluation of connected digit recognition algorithms. It contains 326 informants (111 men, 114 women, 50 boys and 51 girls), each one producing 77 digit sequences. Data are grouped in one training set and one test set. The sequences can include 11 different digits, from "zero" to "nine", plus "oh". The data has been sampled at 20 kHz and digitalized with a resolution of 16 bits.

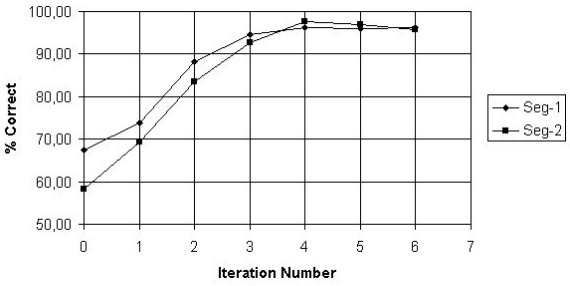

Fig. 2. Evolution of the method for segmentation of the TIDIGITS corpus.

The extension of the database discourages the use of a manual segmentation. For this work paper only the records associated with the adult speakers were processed. The subset of adult training data contains 55 men and 57 women from which specific records were selected for a manual segmentation. The next steps of the method were applied indifferently regarding the assigned set (train or test). On the other hand, the HTK ToolKit [6] was used for the development of the HMMs. The selected vector representation for the speech traces was Mel-frequency cepstral coefficients (MFCC), a 25 ms Hamming window, with an overlap of 5 ms and pre-emphasis coefficient of 0.97 was used. Feature vectors contain 12 MFCC and the log-energy, plus the first and second derivatives. Each vector is composed of 39 real data.

The initial selection of the records for manual labeling was done at random, with two restrictions: first, each of the adult informants in the TIDIGITS training set was represented with one record, and second, this record had to have at least two connected digits. In this way 112 initial records were chosen for manual segmentation and the initial estimation of the reduced models. The labeling of the records was carried out with the SFS [7] tool. To study the effect of the initial manual segmentation two independent segmentation processes were carried out in parallel by two different persons using with the same records.

For both the initial-reduced step and the recursive-extended step 12 models were used, one for each digit plus one for silence. All the models were created using left-right topologies and 12 states for speech plus 2 more for the connections. Figure 2 shows the evolution of the recognition scores at each step of the algorithm recursion produced by the two segmentation processes used. Iteration 0 corresponds to the reduced segmentation and the rest correspond to the recursive process. It can be seen that the initial segmentation is capable of segmenting more than 55% of the records using the reduced HMMs. On the other hand, the recursive iteration produces the segmentation of the majority of the records not segmented initially, and at the same time guarantees an increment of the precision of the initial segmentation.

The evolution of the results shows that the manual segmentation has an influence on the process, however, the final results for both variants barely differ. In the successive iterations the segmentations improve asymptotically and after 6 iterations 96.18 % and 95.72 %, of the corpus records were correctly labeled. An example of a segmentation case is shown in the figure 3.

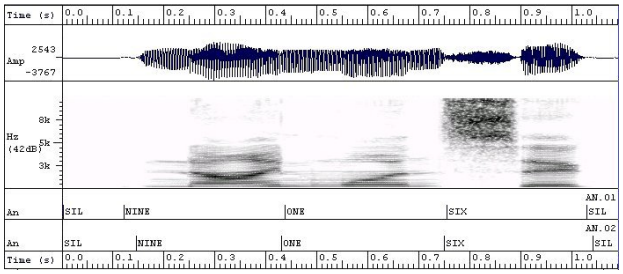

Fig. 3. Example of two extended segmentation for the sequence "nine-one-six".

In the upper part the speech trace is seen. In the middle its broadband spectrogram is shown, and in the lower part the segmentations obtained with both processes are presented. It must be emphasized that this sequence is especially difficult to be manually segmented. However both processes delimit similarly the essentials of digit contours. The differences are larger when the boundary limits are within a pause or silence zone, for example at the beginning and end of the digit sequence. In the essence both variants produce a correct segmentation results, but not identical. The extended segmentation results are

slightly dependent on the initial segmentation by hand.

5 Conclusions

The proposed method can be used for semi-automatic labeling of large speaker-independent databases. It reduces the effects of individual criteria when performing manual segmentation and diminishes significantly the time and cost of the segmentation. The method can be used independently of the purpose of the database, however it is necessary to study its behavior in phoneme segmentation tasks, where the HMMs have to be defined considering contextual information. The proposed method is useful for the development of HMMs in situations where databases are not segmented.

Acknowledgements

This research is being carried out under grant TIC99-0960 from the *Programa Nacional de las Tecnologías de la Información y las Comunicaciones (Spain)*, grant 07T-0001-2000 from the *Plan Regional de Investigación de la Comunidad de Madrid*, and a collaboration contract between *Universidad Politécnica de Madrid* and the *Centre Suisse d'Electronique et de Microtechnique*.

References:

[1] Barras, C., Geoffrois, E., Wu, Z., and Liberman, M., "Transcriber: development and use of a tool for assisting speech corpora production", *Speech Communication*, Vol. 33, No 1-2, January 2001, pp. 5-22.

[2] Cassidy, S., Bird, S., "Querying databases of annotated speech", *Proceedings of the Eleventh Australasian Databases Conference*, IEEE Computer Society, Los Alamitos, CA., 2000, pp. 12-20.

[3] Graff, D., Bird, S., "Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies", *Proceedings of the Second International Conference on Lan-guage Resources and Evaluation*, European Language Resources Association, 2000, pp. 427-433.

[4] Rubio, M., Giménez, V., Gómez, P., "Special Concept Inversion Algorithm for Advanced Cluster Analysis in Self-organizing Maps", *Proc. of the Fifth International Conference on Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies, KES'2001*, September 6-8, 2001, pp. 1180-1184.

[5] Rabiner, L. R. "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, Vol. 77, No 2, February 1989, pp. 257-286.

[6] HTK Toolkit V 3.0. Speech Vision and Robotics Group. Department of Engineering, University of Cambridge http://htk.eng.cam.ac.uk.

[7] Huckvale, M., Speech Filing System (SFS) V. 4.30. University College London, http://www. phon.ucl.ac.uk/