PERFORMANCE EVALUATION TEST OF AN AUTOMATIC SEGMENTATION SYSTEM FOR ITALIAN AND AMERICAN-ENGLISH CONTINUOUS SPEECH.

Fabrizio Balducci, Loredana Cerrato
Fondazione Ugo Bordoni,
Rome, Italy

ABSTRACT

We present in this paper an automatic system for the segmentation[1] of continuos speech. The segmentation algorithm has been implemented modifying the CDHMM system RES (Recognition Experimental System) for the automatic recognition of continuos speech. The segmentation task is performed as a phonetic recognition constrained by the grammar model built only on the exact sequence of uttered phonemes.

We also report the results of some performance evaluation test using TIMIT (American-English) and APASCI (Italian) database. In order to evaluate the system performance also on telephonic application, we modified the Italian database to produce a telephonic version. Therefore we report both the results for normal and telephonic speech focusing our attention on the performance of vowel segmentation. As our envisaged application is the segmentation of speech recorded by means of wire-tapping for forensic application, we also report some specific results for the identification of phonemic boundaries on a real wire tapped speech.

1. INTRODUCTION

Automatic labelling of speech signals is an increasingly important problem in many applications in speech technology where large amount of segmented and labelled speech data is required (for instance to train and test ASR systems).

Alignment of speech and text is usually manually performed by phoneticians, but this activity is extremely time consuming and costly. Moreover, notwithstanding the general criteria for segmentation and labelling, due to human variability of visual and perceptual capabilities, there is always disagreement among skilled human labelling experts in their results of labelling the same waveform.

In order to speed up the task and to avoid the disagreement, researchers are trying to develop automatic labelling methods, which of course cannot be so reliable as human performances, but which at least appear to be faster, cheaper and when they make errors they always make them in a coherent way [1].

The system we present is not intended to replace the skilled phonetician, but aims at providing a support to the task of labelling, in particular for application in the field of speaker identification for forensic purposes, where it is often necessary to segment big amount of recorded speech.

In our lab speaker identification and verification is performed by means of the IDEM [2] method, which is based on the comparison of a set of parameters (the frequency of F0 and the frequency of the first three formants of the Italian vowels /a/, /e/, /i/, /o/) estimated in stable portion of speech and on a rigorous statistical procedure to estimate false rejection and false identification probabilities. In order to extract the formant values is often necessary to segment speech, often recorded over the telephonic line by means of wire-tapping; therefore we thought it could be useful to implement an automatic system, particularly reliable in the identification of vocalic segments in telephonic signals, even in the presence of noise.

2. THE SEGMENTATION ALGORITHM

The system we describe automatically carries out the segmentation of the speech sample starting from its phonetic transcription. It is supported by an automatic transcription system, based on a series of phonological rules able to perform a grapheme-to-phoneme conversion of the input string (in SAMPA conventions). The rules are entirely based on the orthographic shapes of the Italian words and they do not take in consideration any level of grammatical knowledge[3].

The segmentation algorithm has been realised modifying the automatic recognition system RES [4] for continuous speech based on the use of hidden models of Markov (HMM) for the probabilistic modelling of the acoustic events. Such technique is based on the supposition that the vocal production happens through succession of acoustically stationary events corresponding to articulatory configurations of the various phones [5].

Each phone is identified by an HMM model composed of a sequence of N_s+2 states s=0,1,…, N_s where the first and the last states are conventional non-emitting states (i.e.with no associated observation). Such states are connected by arrows in a left-to-right structure to represent the possible successions of acoustic events. The model topology includes also recursive transitions upon states to confirm the same state for more time steps (the so-called analysis step), this is to allow the best alignment, in probabilistic sense, of phonemes models to the vocal sequence to recognise.

A distribution probability function is associated to every emitting state and it is represented by multivariate Gaussian mixtures:

[1] We use either segmentation or labeling in the same sense of aligning a waveform to a phonetic transcription

$$P(\mathbf{y} \mid s = i) = \sum_{k=1}^{M} c_k^{(i)} N(\mathbf{y}, \mu_k^{(i)}, \mathbf{U}_k^{(i)})$$

where $\mu_k^{(i)}$ are mean vectors of the covariance matrices $U_k^{(i)}$ and $c_k^{(i)}$ are weights of the Gaussian pdf *N(.,.,.)* with $\sum_{k=1}^{M} c_k^{(i)} = 1$ for consistency.

Each Markovian model is so tuned on a phone, hence a good segmentation depends by an accurate training step of the acoustic models. In RES the acoustic training is based on Baum-Welch procedure to estimate the pdf parameters; the number of gaussians for each mixture is fixed during the training step according to the quantity and quality of available speech sounds.
On the contrary the topology model, that is the number of emitting states and the allowed transitions, is a priori fixed.
During the experiments the best results were obtained with a 6 state-model for Italian and 3 state-model for American English, both with transition between consecutive states and between the i-th and i+2th. Finally it is important to note that even if the choice of the start model can influence the final performance, it doesn't constrain the training procedure to keep the less probable transition states.
The Viterbi algorithm [6], implemented in RES for recognition tasks, has been modified to keep trace of the index of speech segments (frames) corresponding to phoneme transitions.
In the segmentation the idea is to build a composite global HMM over the utterance, by concatenating single phonetic models of phonetic transcription; so with the Viterbi algorithm it is possible to find the state sequence that is most likely to have generated the observation and then know how many observations are associated to each state in the optimal final sequence. From this information it is possible to assign phonetic boundaries with a time accuracy almost equal to the analysis step.
To perform the segmentation it is then necessary to supply, beyond the acoustic signal, also the information about the phonetic transcription of the processing signal, the Figure 1 shows the system data flow:

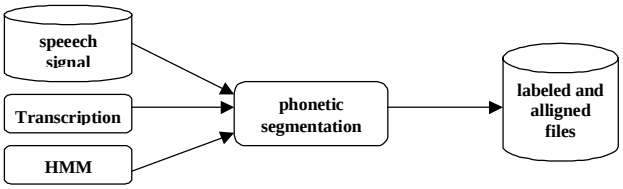

Figure 1 *Data flow of the segmentation system.*

As concerning the pause between consecutive words, the system inserts them as options in the composite HMM model and segments them phonetically only if they are present in the acoustic signal.
Figure 2 shows an example for the words couple "ogni anno", the double arrow after the phoneme *i* allows to the Viterbi algorithm to find the best phoneme (*sil* or *a*) that match the acoustic representation

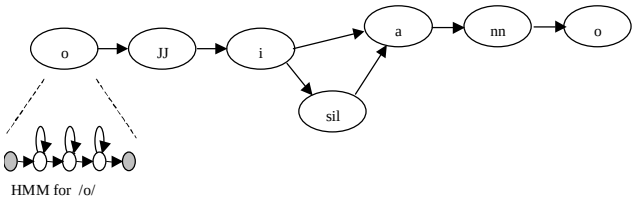

Figure 2 *Schematic representation of words concatenation.*

For an algorithm optimisation the "*beam search"* technique [7,8] is applied to dynamically prune the hypothesis produced at every step.

3. EVALUATION METHOD

The qualitative evaluation of the reliability of our procedure of automatic segmentation has been carried out verifying the consistence of the boundaries proposed by the automatic procedure with those given in the reference segmentation. Therefore the errors of the automatic procedure are evaluated in function of the deviation from the corresponding reference segmentation.
Two types of errors have been defined:

- ***position errors***, characterised by the fact that the phoneme is correctly identified but there isn't an absolute overlapping with the correspondent boundaries and
- ***recognition errors***, characterised by the fact that the phoneme is indicated out of his existence area.

While for the first case we carry out a quantitative evaluation, we consider the second case as incorrect. Figure 3. clarifies the situations.

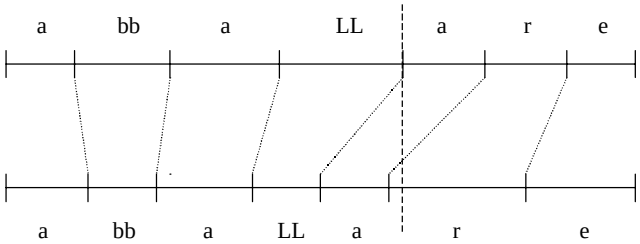

Figure *3. Example of possible errors: the automatic segmentation produces a* ***recognition error*** *for the second realisation of /a/ and* ***position error*** *for the other phonemes.*

Recognition errors correspond to critical situation for the segmentation and can also cause errors on the segmentation of following phonemes. For this reason it is extremely important to limit their number.
Position errors are less critical, and it is possible to classify as "correspondence" also for deviations which do not exceed 5 ms.
Other results [9] fix to 10 ms the threshold under which it is not necessary to manually correct the position errors. To give a more accurate description of system performance and following the consideration reported in [10] we also take into account results for deviations included below 20 ms.

4. EXPERIMENTAL RESULTS

Our test have been performed using the Italian speech database

APASCI [11] and American English database TIMIT [12]. Both databases are similar in their structure: they consist of production of read utterances divided so to allow the development and testing of speech recognition systems both speaker dependent and speaker independent; they are provided with a reference segmentation for the evaluation of the results.
The TIMIT database was used to compare our system performance with other similar systems.
Being telephonic speech most suitable for our envisaged forensic application, we transformed the APASCI signal into telephonic quality, by means of an artificial mouth and a telephone-line simulator.
The best results turned out to be obtained using, for the extraction of the characteristics of the acoustic signal, 9 cepstral coefficients, the energy and their derivatives of the first and second order; the number of Gaussian for every varied mixture from phoneme to phoneme until a maximum of 12 according to the results of the training. For TIMIT database we used the same configuration parameters, but the cepstral coefficient were 12.

4.1 TIMIT results

The speech material used for the experiments is made of read utterances produced by 438 male speakers divided as follow:

326 speakers for the acoustic model training
112 for the test

We use a reference phonological model made of a subset of 39 units from the 54 unit TIMIT phonemic set.
The number of emitting states of the model is equal to 3.
Segmentation results for Timit male speakers are reported in Figure 4.

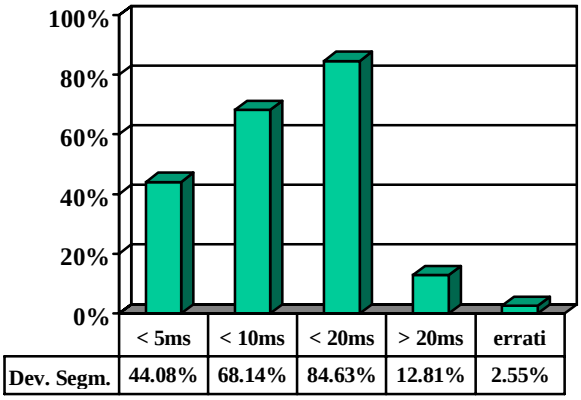

Figure 4 *Results for TIMIT male speakers only.*

4.2 APASCI results

The speech material used for the experiments is made of read utterances produced by 106 male speakers divided as follow:

- 50 male speakers form the SI (speaker independent) group:
  30 speakers for the acoustic model training
  20 for the test.
- 56 female speakers form the SI group:
  36 speakers for the acoustic model training
  20 for the test.

For Italian we refer to a phonological model made of 30 units[13] with two more units: one for the silence and one for shwa
The number of emitting states of the model is equal to 6. The experiments were carried out training the acoustic models on the indicated portion of the database referred to as training set inside the documentation of the corpus. The test was carried out on the remaining files corresponding to the speakers which were not used for the training. The results relative to position errors and recognition errors for male and female speakers are shown in Figure 5.

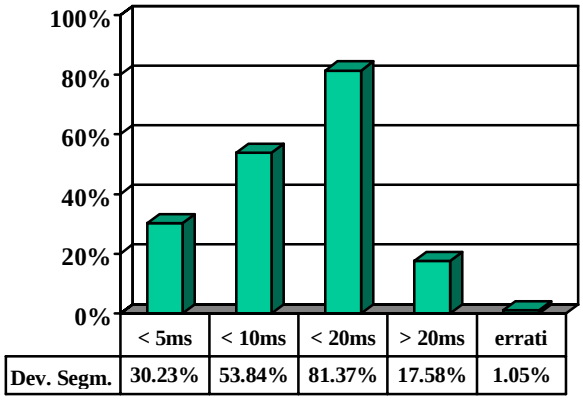

Figure 5 *Final results for APASCI male and female speakers*

We also trained the model using only the male speakers and then we carried out the segmentation on their speech; this way we obtained an increasing of the performance as shown in Figure 6.

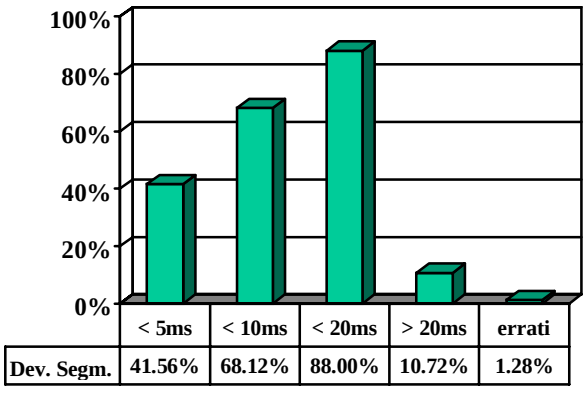

Figure 6 R*esults for male speakers only*

Due to our particular interest in speaker identification and to the need of extracting formant values form vocalic segments, we evaluated the position errors for the vocalic phonemes; the results are reported in Figure 7.
The use of formant data in the forensic framework is well recognised to give a fairly consistent measure of speaker identity, even in the presence of noise. This is mainly due to the higher signal to noise ratio in spectral peak regions when compared to other parts of the spectrum, and to the strict

relationship between formant peaks, and peaks dynamics, to the physical behaviour and personal attitudes of the speaker.

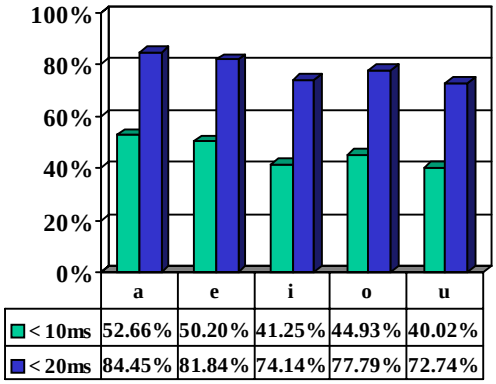

Figure 7 *Position error for vowels*

The system also allows an analysis of phonemic duration, for instance in Tab. 1 we report the mean duration of all the vocalic sounds identified in the APASCI database.

a	e	i	O	U
101.30ms	87.65ms	82.59ms	92.83ms	93.15ms

Tab. 1 *Mean duration of all the vocalic sound in APASCI*

4.3 Other results

The system has been also tested using some speech material extracted from real wire tapping and the initial obtained results are very promising: even if the performance is not optimal for all the phonemic boundaries, it appears to be good enough for the identification of vocalic segments.

In Figure 8 we report a segmentation example on a fragment of real wire tapping speech, using the software Snack[www.speech.kth.se].

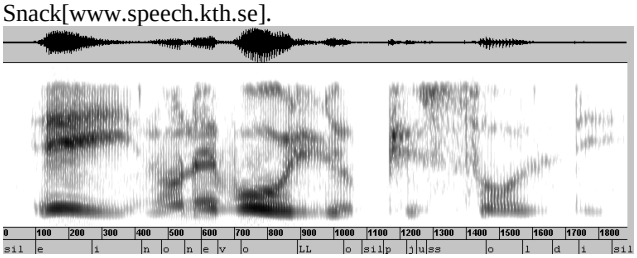

Figure 8 *Example of segmentation on wire tapping speech for the utterance "ei non ne voglio più soldi".*

5. CONCLUSIONS

The results obtained with TIMIT database under 5ms (44.08%) and 10ms (84.63%) show the reliability of the performance of our system and are consistent with those obtained by other similar systems [14] [15]. The telephonic version of APASCI database was useful to train the Italian acoustic models and to test system performance in adverse environment conditions. Being the results quite promising (under 5ms: 41.56%, under 20ms: 88.00%) we started to use the system in real application, to segment speech recorded by means of wire tapping. The system showed to be quite reliable in segmenting vocalic items, which are very useful in our speaker identification procedure.

REFERENCES

[1] Cosi P., Falavigna D. and Omologo M., *"A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies*", Proceedings of EUROSPEECH-91.

[2] Paoloni A.: "*Text Indipendent Speaker IDEntification Method: IDEM*", American Academy of Forensic Sciences Meeting, Colorado, February 1999.

[3] Cerrato L., D'Alterio D., Di Carlo A., "*Regole di trascrizione da grafema a fonema per l'Italiano standard. del XXVII conevgno AIA Genova"*, 1999, in press.

[4] Prina Ricotti L., Becchetti C., *"Speech Recognition, Theory and C++ Implementation"*, Wiley, 1998.

[5] Rabiner L., Biing-Hwang J., *"Fundamentals of speech recognition"*, Prentice Hall, N.J., 1993

[6]Viterbi A.J., *"Error bounds for convolutional codes and asymptotically optimum decoding algorithm"*, IEEE Transaction on Information Theory 13,1967.

[7]Lowerre B., Reddy R., *"The harpy speech understanding system"*, in W.A.Lea Ed., Trends in Speech Recognition, Prentice Hall, N.J., 1980

[8] Bridle J.S., Brown M.D., Chamberlain R.M., *"An algorithm for connected word recognition"*, Proc. of ICASSP 1982, Paris;

[9] Barry W.J., "SALA labelling test", Summary report, SAM-document: SAM UCL-020, 1991.

[10] Erp A. Van and Boves L., *"Manual segmentation and labelling of speech"*, Proc. Speech 1988 Edinburgh , Scotland, p.1131-1138

[11] Angelini 94 Angelini B., Brugnara F., Falavigna D., Giuliani D., Gretter R., M. Omologo M., *"Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus"*. In Proceedings of ICLSP, Vol. III, Yokohama, Japan, 1994, pp.1392-1394,.

[12] Lamel L., Kassel R. and Seneff S., *"Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus"*, Proceedings of the DARPA Speech Recognition Workshop, pp.26-32, March 1987.

[13] Lepschy A.L., Lepschy G.C, 1981, La Lingua Italiana, Bompiani, Milano.

[14] Kvale K., "*Segmentation and labelling of speech*", Universitetet I Trondheim, '93.

[15] Ljolje A., Riley M.D., "*Automatic segmentation and labelling of speech*", ICASSP '91, pagg. 473-476.