UNINA System for the EVALITA 2011 Forced Alignment Task

Bogdan Ludusan*

LUSI-lab, Università degli Studi di Napoli "Federico II", Naples, Italy
ludusan@na.infn.it

Abstract. This report presents the system proposed and the results obtained by our group for the EVALITA 2011 Forced Alignment on Spontaneous Speech task. The system is composed of a module for acoustic modelling, which uses the SPRAAK toolkit, and a second one for the processing of textual information. Several tests were performed to determine the impact of frame shift size and speaker adaptation on the accuracy of the alignment. Good segmentation results were obtained, the proposed system outperforming the other teams' systems, but performance improvements can be achieved by using a pronunciation dictionary.

Keywords: Hidden Markov Models, Forced Alignment, Spontaneous Speech

1 Introduction

Forced alignment is an invaluable technique nowadays with the existence of very large speech corpora and the need of fast and inexpensive tools for their segmentation. Since manual segmentation is both time consuming and financially costly the use of automatic tools, performing forced alignment, is preferred. Viterbi alignment is used extensively in speech research for different topics, ranging from speech recognition (e.g. [1]), to speech synthesis (e.g. [2]) and phonetic analysis (e.g. [3]).

While for other languages alignment tools were evaluated on spontaneous speech (e.g. [3], [4]), for Italian, most of the papers reporting forced alignment results used either read or laboratory speech for the evaluation (e.g. [5], [6]). And although good results were presented in these papers, around 90% of phoneme boundaries placed within 20 ms [5] and 94% respectively [6] on a corpus of read speech, we lack an evaluation of this tools on conversational speech.

This paper is organized as follows: first, the task for which the system was conceived is introduced in section 2, then the system itself is presented in section 3. Section 4 illustrates the results obtained on both subtasks, while the paper is concluded with a discussion of the results and some conclusions are reached.

* The author is currently affiliated with CNRS-IRISA, Rennes.

2 Task Definition

EVALITA is the Italian evaluation campaign for spoken and written language and one of the speech tasks proposed in the current campaign is forced alignment. The aim of this task is to automatically align a speech sequence with its corresponding orthographic transcription. Two subtasks were proposed: one for phonemic alignment, while the second one for word alignment.

Dialogues from a freely available corpus of spoken Italian, CLIPS [7], was used both for training and testing. The training set consisted of 15 map task and difference test dialogues, each one having a duration of 7-20 minutes. The audio files were distributed together with their corresponding phone-level and word-level transcriptions. For testing, a 10 minutes subset of CLIPS containing the same types of dialogues was chosen and the canonical form of the uttered words provided.

3 System Presentation

The system used for the forced alignment task is based on the standard architecture (see block scheme in Figure 1). Its central part is the Viterbi aligner and it also has some modules for the processing of textual information. The alignment function, as well as the acoustic model training components are part of SPRAAK [8], an open source speech recognition toolkit. As it can be observed from the system block scheme, the same procedure is followed for both subtask with only one extra step for the word segmentation subtask.

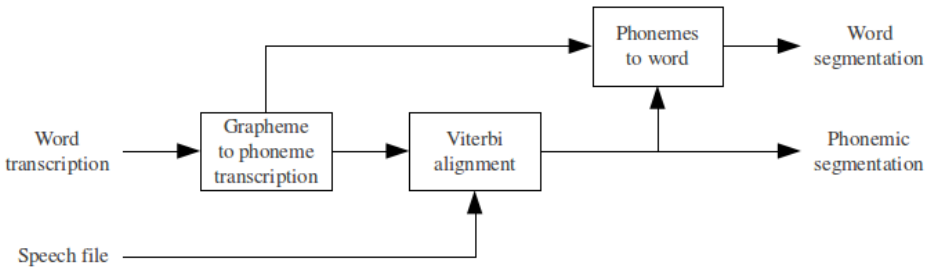

Fig. 1. The block diagram of the system.

3.1 Data processing

For the training step no additional data preparation was needed, except for the conversion of the input data to the format used by SPRAAK. But in order to obtain the phone-level transcription of the test data, a grapheme-to-phoneme conversion software had to be used. WR2ST [9] takes in input a text sequence

and outputs its corresponding phonological transcription. Please note that only the canonical form of the word is given, without any other pronunciation variants.

The phonemic transcription obtained after the previous step is employed in two subsequent steps. First, it is given as input to the Viterbi alignment function along with the speech file it refers to. And second, it plays a role in obtaining the word segmentation, being used to align the phoneme-level segmentation with the transcribed sequence of words. The later is done by means of a script which, given the orthographic transcription of a file, the phonemic transcription of each word contained in it and the corresponding phoneme boundaries, returns the word segmentation.

3.2 Acoustic models

Context independent (CI) phone models were trained for the forced alignment task, one for each symbol present in the phn inventory set supplied with the training data. It consists of: all the Italian vowels (7), all the Italian consonants except the semivowels /j/ and /w/ (21), several diphthongs (13), a silence model and a garbage model. The silence model corresponds to the silent pauses present in the train set, while the filled pauses and undesired phenomena were modelled into the garbage model.

Both the training procedure and the Viterbi aligner use first a preprocessing step in which frame-based features are extracted. The spectral features are extracted from a 32 ms Hamming analysis window and file-based mean normalization is applied. After taking the first and second order derivatives, a 72-dimensional feature vector is obtained for each frame. Features are extracted on a frame based. During the training process, a Mutual Information Discriminant Analysis (MIDA) is performed on the features space to reduce its dimensionality. As a result, only the most informative 36 spectral coefficients are kept.

SPRAAK uses an acoustic modelling technique called semi-continuous HMM (SCHMM), in which HMM states are globally tied. This means that the pool of Gaussians is shared by all the states, a state being distinguished from another one only by the weights of the Gaussians in the mixture. The HMM models trained have 3 states and a left-to-right topology.

The training procedure is performed in two steps. The first step consists itself of three processes: MIDA training, initialization of tied Gaussian mixtures, followed by three iterations of Viterbi training. The second step performs a Viterbi training pass preceded and followed by full covariance estimation and decorrelation passes.

4 Results

Four different models were trained and tested, by varying two parameters: the frame shift between to consecutive analysis windows and whether the models use or not speaker adaptation. Two values of the frame shift were tested: 10 ms, the standard value used in speech recognition systems, and 5 ms, which

was used in some studies with the intent of obtaining more accurate segment boundaries (e.g. [3]). The speaker adaptation method chosen was the vocal tract length normalization (VTLN). It aims to normalize the spectral variations due to the different tract lengths between male and female speakers by warping the spectrum of a speaker towards a global average vocal tract length.

For the evaluation of the results the NIST sclite package [10] was employed. The method used, the so called time-mediated alignment is a variant of dynamic programming alignment, which minimizes a distance function between pairs of words. The particularity of this method is the fact that it computes the word-to-word distance based on beginning and ending word times.

The results obtained for the phone-level segmentation are presented in Table 1. A Wilcoxon matched-pairs signed-rank test was performed in order to test the statistical significance of the results obtained. It showed a significant difference at the 1% level between the two runs using a 5 ms frame shift and the (10 ms, no VTLN) run. When comparing against the (10 ms, VTLN) run, a significant difference at the 5% level was found with respect to the (5 ms, no VTLN) run, but no significant difference with the (5 ms, VTLN) run.

Table 1. Results obtained on the phoneme alignment subtask.

frame shift [msec]	VTLN	Substitutions [%]	Deletions [%]	Insertions [%]	Errors [%]
5	no	5.0	2.0	8.1	15.1
5	yes	5.2	1.8	8.2	15.1
10	no	4.9	1.2	7.2	13.3
10	yes	5.1	1.3	7.2	13.6

Table 2 illustrates the system results on the word segmentation subtask. There seems to be no significant difference between the results obtained with different acoustic models.

Table 2. Results obtained on the word alignment subtask.

frame shift [msec]	VTLN	Substitutions [%]	Deletions [%]	Insertions [%]	Errors [%]
5	no	0.1	0.5	0.5	1.2
5	yes	0.2	0.8	0.8	1.8
10	no	0.2	0.6	0.6	1.4
10	yes	0.2	0.5	0.5	1.2

5 Discussion

The use of speaker adaptation appear to have little impact on the alignment accuracy, as the models using VTLN show similar performance compared to the ones without adaptation. This finding is consistent with the results of a study that evaluated the influence of several factors on the forced alignment accuracy [11]. Regarding the frame length, the use of a smaller frame shift decreases the segmentation accuracy. While this performance drop can be partially explained by a higher incidence of very short (less than 30 ms) flasely detected phones at the edge of garbage regions, a more in depth examination of the results is needed for a complete explanation.

The first step taken in the analysis of the errors produced by the alignment procedure was to determine how precise are the boundaries of the segments found to be correct. Table 3 shows the shift of the automatically placed boundary with respect to the reference for both phone- and word-level alignments. One can see that almost 97% of the phonetic markers are positioned within 40 ms of the manual boundary. By comparing the boundary shift obtained for the two alignment levels, it can be concluded that the intra-word segment boundaries are more precise than the inter-word boundaries. This may be due to the poor modelling ability of the garbage class, which contains filled pauses and word lengthening.

Table 3. Boundary marker shift from the reference.

Level	$\leq$ 10ms [%]	$\leq$ 20ms [%]	$\leq$ 30ms [%]	$\leq$ 40ms [%]	$>$ 40ms [%]
Phone	56.59	82.20	93.78	96.86	100
Word	46.04	70.46	86.03	91.02	100

A part of the errors produced are due to the lack of precision in the placement of the boundaries, but a significant amount of them are due to the phenomenons occurring in connected speech (like reductions) which change the pronunciation of words. In our case, one of the most important source of errors is the fact that the evaluation process considers, in the case of two adjacent vowels, also the possibility of the two forming a diphthong. And because the canonical form of the words contains always two vowels, that translates into a substitution and an insertion error each time. An analysis of this phenomenon revealed that 40% of the total number of substitutions and insertions is due to it. Other types of errors can be traced to the deletion centralization of vowels.

While there is no possibility of improving the performance of the system without taking into account different pronunciation variants for each word, the current system accuracy can be considered as an upper bound of the error.

Using additional sources of pronunciation would only increase the aligner's performance.

6 Conclusions

This paper described a system for the forced alignment of speech files, at the phone- and word-level. the system was tested on the a corpus of Italian spontaneous speech, with good results. Because the current approach does not use any pronunciation variants for the words in the lexicon, it is not able to recover from errors due to words pronunciations different from their canonical form. The next step in improving the system would be the addition of a pronunciation dictionary as well as providing it with a set of phonetic rules for the changes that occur during conversational speech.

Acknowledgements. The author would like to thank Dino Seppi for providing the gender models for the VTLN.

References

1. Wu, S.L., Kingsbury, B., Morgan, N., Greenberg, S.: Incorporating information from syllable-length time scales into automatic speech recognition. In: Proc. of IEEE ICASSP, pp. 721–724 (1998)
2. Kominek, J., Bennet, C., Black, A.W.: Evaluating and correcting phoneme segmentation for unit slection synthesis. In: Proc. of EUROSPEECH-2003, pp. 313–316 (2003)
3. Schuppler, B., Ernestus, M., Scharenborg, O., Boves, L.: Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. Journal of Phonetics 39, 96–109 (2011)
4. Greenberg, S., Chang, S.: Linguistic dissection of switchboard-corpus automatic speech recognition systems. In: Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, pp. 195–202, (2000)
5. Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., Omologo, M.: Automatic segmentation and labeling of English and Italian speech databases. In: Proc. of EUROSPEECH'93, pp. 653–656, (1993)
6. Cangemi, F., Cutugno, F., Ludusan, B., Seppi, D., Van Compernolle, D.: ASSI - automatic speech segmentation for Italian: tools, models, evaluation and applications. In: Proc. of the 7th AISV Conference, (2011)
7. Savy, R., Cutugno, F.: CLIPS: diatopic, diamesic and diaphasic variations of spoken Italian. In: Proc. of the 5th Corpus Linguistics Conference (2009)
8. Demuynck, K., Roelens, J., Van Compernolle, D., Wambacq, P.: SPRAAK: an open source SPeech Recognition and Automatic Annotation Kit. In: Proc. of INTERSPEECH-2008, pp. 495–499. (2008)
9. WR2ST, `http://www.clips.unina.it/`
10. Sclite software package, `http://www.nist.gov/speech/tools/`
11. Chen, L., Liu, Y., Harper, M., Maia, E., McRoy, S.: Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus. In: Proc. of Language Resource and Evaluation Conference, pp. 759–762 (2004)