SAD-Based Italian Forced Alignment Strategies

Giulio Paci, Giacomo Sommavilla, and Piero Cosi

Istituto di Scienze e Tecnologie della Cognizione (ISTC) - Sede di Padova,
via Martiri della Libertà, 2, 35137 Padova, Italia
{giulio.paci,giacomo.sommavilla,piero.cosi}@pd.istc.cnr.it
http://www.pd.istc.cnr.it

Abstract. The Evalita 2011 contest proposed two forced alignment tasks, word and phone segmentation, and two modalities, "open" and "closed". A system for each combination of task and modality has been proposed and submitted for evaluation. Direct use of Silence/Activity detection in forced alignment has been tested. Positive effects were shown in the acoustic model training step, especially when dealing with long pauses. The exploitation of multiple forced alignment systems through a voting procedure has also been tested.

Keywords: Evalita 2011, Italian, Forced Alignment, SAD, Acoustic Model training, Voting strategy, Data Preparation.

1 Introduction

Forced alignment refers to the problem of automatically aligning speech audio and its transcription, by identifying and properly marking time ranges in the speech data corresponding to each unit (words, phones, ...) in the transcription.

Forced alignment is of interest for all those applications where a large amount of time aligned labelled data is required, such as automatic speech recognition, unit-selection [4] and HMM-based statistical parametric speech synthesis [12,8] or speech-based linguistic studies in general. In the speech synthesis domain, usually audio data comes from a single speaker and forced alignment procedures take advantage of that. In the general case, forced alignment should be carried on multiple, unknown, speakers.

Before the Evalita 2011 campaign, only a few studies reported evaluation results for speaker independent forced alignment for Italian spontaneous speech. The Evalita 2011 evaluation campaign proposed two forced alignment tasks for the Italian language: the word forced alignment and the phone forced alignment. Two different modalities have been allowed for each task, open (OM) and closed (CM). The Evalita 2011 campaign showed that the performance of current forced alignment systems for Italian is similar to those reported by the state of the art systems for other languages.

This paper describes the development of the systems proposed by the authors for both tasks. All the systems have been implemented using SONIC [9], the University of Colorado large vocabulary continuous speech recognition system,

B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 322–329, 2013.

and AudioSeg [7], the INRIA audio segmentation and classification toolkit. Three systems were tested for the word forced alignment task and the final system is based on a voting procedure among them. This system scored an overall accuracy of 97.4% (OM) and 98.4% (CM). The best performing of the three systems tested for the word forced alignment task has been proposed for the phone forced alignment task. This system scored 90.6% (OM) and 92.4% (CM).

A previously developed acoustic model has been used for the OM. Within this work, results reported for the OM represents the baseline, as target-specific training data has been used only for tuning. For the CM systems an acoustic model trained solely with the Evalita 2011 training set has been used. We tested the same training procedure used to obtain the OM acoustic model and a slightly different one involving the direct use of a Silence/Activity Detection (SAD) algorithm. A description of both procedures and their comparison is provided in Section 4.2.

2 Data Description

2.1 Evalita 2011 Training Data

Evalita 2011 training data is a subset of the CLIPS corpus [10], consisting of about 5 hours of spontaneous speech from 90 adult speakers from different Italian areas, collected during map-task experiments. The corpus contains aligned orthographic and phonetic transcriptions for all the recordings. The audio recordings are clean, although during long pauses low energy non-transcribed speech is intelligible. Some errors has been found in both the orthographic and the phonetic transcriptions. In Section 3 a description of the major errors and actions taken to cope with them is reported.

2.2 Development Set

For internal evaluation a development test set has been extracted from the Evalita 2011 training data. The corpus has been divided by speaker and sorted by increasing size. Four speakers, two males and two females, have been selected from the top of the list, so that the development test included at least one northern and one southern Italian accent speaker.

3 Data Preparation

Phonetic transcriptions in the Evalita 2011 training data contain many garbage fillers, even when intelligible phones can be heard in the corresponding audio file. For this reason it has been decided to avoid its direct use in the training phase. Instead it was used to obtain a customised phonetic lexicon to be used in the training step. A previously developed phonetic lexicon and a letter to sound module were used to provide possible phonetic transcriptions. The phonetic dictionary was augmented using entries (including mispronounced words) derived from the alignment of .WRD and .PHN files.

Problems arose with words that end with a vowel, followed by a word that also starts with a vowel. In such cases, typically, a diphthong is found on the border between two stop/start word marker, aligned in the boundary between the two words. For example, in the phrase "`gli occhiali`", the .PHN file reports a "`jo`" phoneme belonging to both words. In these cases we converted the shared phoneme into two distinct phonemes, each one belonging to a different word, taking the stop/start word marker as dividing time instant.

Further problems arose in some words/phones misalignment cases (see Fig. 1). In such cases, in order to create the dictionary, we followed this rule: if a word ends (starts) with a letter, but it is aligned with a phoneme that is not related to that letter in the .PHN transcription, and that phoneme is related with the first (last) letter of the next (previous) word in the transcript, we systematically deleted that phoneme from the word phonetic transcription.

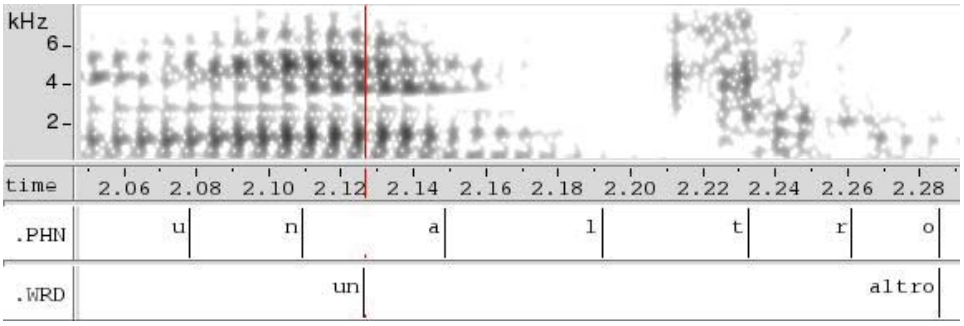

Fig. 1. Example of phonemes' mismatch due to alignment: the word "`un`" has been incorrectly phonetized "`u n a`"

The last problem we faced was dealing with garbage fillers. If a word transcription contained a single garbage filler, we manually replaced it with tokens derived from the standard pronunciation of the words (e.g., the phonetic transcription "`a o r *`" of the word "`allora`", has been changed into "`a o r a`"). If a word transcription contained more than one filler, we simply discharged it.

There were also cases in which the orthographic form of a word was just a garbage filler symbol (e.g., `*` or `#`), and the phonetization was made up of several non-garbage phonemes, out of which intelligible words could be recognised by manual corrections. We set up a procedure for retrieving all those garbage-words whose phonetizations consisted of more than three phonemes. Those words and related orthographic transcriptions have been manually corrected.

Finally we listed all the words with non alphabetic characters we found in the Evalita 2011 training set. We checked by hand those words: some of them were typos (especially with filler words) and have been manually fixed. This was an important step because it helped us avoiding errors derived from the letter to sound model trying to provide phonetic transcription of misspelled fillers (e.g., "`<inspiration`" instead of "`<inspiration>`").

4 Forced Alignment Procedures

Each proposed system makes use of several resources: a Silence/Activity Detector (SAD), an acoustic model, a phonetization module and a speech recognition engine. In this section we will briefly describe each resource and how it is used in the proposed systems.

4.1 Silence/Activity Detection

The use of an energy-based SAD has been introduced to cope with long pauses filled with low-energy non-transcribed speech that occur often in the Evalita 2011 training data and may confuse automatic systems. The algorithm employed for SAD [7] operates in two steps. In the first step it estimates a bi-Gaussian model of the log-energy of audio frames. Using this model and the maximum likelihood criterion, an optimal threshold is estimated. In the second step, frames are classified into silence or activity according to the energy threshold and to a constraint imposed to the minimum duration of silence sequences. In all the experiments a frame length of 200ms has been used. Silence sequences shorter than 100ms has been ignored, and thus considered "activity". Finally a margin of 50ms has been added around all the activity sequences.

4.2 Acoustic Models

The acoustic model trainer for SONIC is based on sequential estimation using Viterbi forced alignment and phonetic decision tree state clustering. Alignments were initially boot-strapped using a default U.S. English acoustic model (adult 16 kHz microphone speech), as described in [5]. The bootstrapping procedure consists of the following steps: (1) each of the 40 Italian SAMPA phonemes is mapped into an acoustically similar U.S. English one; (2) this phonetic mapping is then used (along with an U.S. English acoustic model) to align the training data; (3) a first acoustic model for an Italian phoneme set is built with those alignments using decision tree state clustering; (4) finally, the alignment/training procedure is repeated several times to obtain improved alignments and model parameter estimates.

Our acoustic models consist of gender-independent triphones using standard 39-dimensional PMVDR features that have proven [11] to be more effective than traditional MFCC features: (1) they are robust in noisy environments, which typically affects low-energy portions of the spectrum, because they model well the peaks (the "upper" region) of the spectrum; (2) they incorporate perceptual considerations, i.e., the power spectrum peaks are computed at data-dependent warped frequencies, thus leading to an envelope modelling approach that is more appropriate than fixed-filterbank spectral ones for a broad range of speech phone classes.

Open Modality Acoustic Model. In the Open Modality (OM) any type of data was allowed for system training, including the provided training data. In

this case the acoustic model was built upon the APASCI corpus only, which consists of about 10 hours of read speech from 100 adult speakers with both orthographic and phonetic transcriptions [1,2]. The Evalita 2011 training set has been used solely to assess the reliability of the tested systems.

Closed Modality Acoustic Model. In the Closed Modality (CM), after the data preparation step, we proceeded with SONIC training procedure, using only Evalita 2011 training data (orthographic transcriptions and audio files) and a phonetic dictionary, that has been built as described in Section 3. We didn't use the provided phonetic information in .PHN files because of missing phonemes in the transcriptions.

We trained two different acoustic models. The first system has been built up by audio files that have been processed by the energy based SAD described in Section 4.1. We avoided the use of timing information of .WRD files in order to demonstrate the effectiveness of the procedure when only the orthographic transcription is available. The second acoustic model was trained without SAD information. We tested the two systems in a word forced alignment task. As shown in Table 1, it turned out that the first system worked a little better than the second, thus helping to confirm our impression that background voice could badly train silence models.

Table 1. Acoustic Models and Strategies Comparison (word forced alignment)

System	OM_{train} (%)	OM (%)	$CM_{AM\ std}$(%)	$CM_{AM\ SAD}$(%)
$SONIC_{base}$	**97.0**	**97.3**	97.2	**98.2**
$SONIC_{del}$	96.7	97.1	**97.5**	98.0
$SONIC_{SAD}$	95.8	96.5	96.6	96.7
Voting	**97.2**	**97.7**	**97.6**	**98.3**

4.3 Phonetization Module

Neither the word forced alignment nor the phone forced alignment tasks of Evalita 2011 assume the availability of phonetic transcriptions as one of the input of the aligners, thus a phonetization module is required in order to hypothesise it. This is especially true for the phone forced alignment task, where the task depends on a phoneme recognition subtask.

In our systems the phonetization module provides phonetic transcriptions for each word in the orthographic transcription by first looking into a phonetic lexicon and then employing a decision tree-based letter to sound algorithm [9,3] for missing words, implemented by SONIC. The letter to sound training procedure comprehends two steps: (1) firstly, an alignment between the letters in an entry and the phonemes in its pronunciation is automatically computed (letters can map to zero, one, two or very exceptionally three phonemes); (2) once alignment is completed, the phoneme prediction model is trained: for each letter in the alphabet, a CART tree has been built, given the letter context (the three

preceding and the three following letters) to predict zero, one or more phonemes from the aligned data.

For this work, the decision tree was trained upon an Italian phonetic lexicon of about 500k Italian forms, originally developed for speech synthesis [6] and then adapted for speech recognition: common alternative transcriptions of some words have been added, gemination and syllable division information has been discharged. For the CM the stress marks have been discharged as well.

4.4 Word Alignment

The system used for the word forced alignment task is based on a voting procedure, described at the end of this section, among the following three subsystems.

$\text{SONIC}_{\text{base}}$**:** this is our baseline system, made up by the SONIC aligner with its integrated Voice Activity Detector (VAD);

$\text{SONIC}_{\text{del}}$**:** this is identical to the baseline system, but the aligner is allowed to discharge phonemes from the transcriptions if their probability is low;

$\text{SONIC}_{\text{SAD}}$**:** this is the SONIC aligner using an external SAD front-end and with the integrated VAD disabled. Following the same intuition behind the SAD-based training procedure explained in Section 4.2, we tried to filter out low energy (non-transcribed) speech prior to perform the alignment. This requires silence reintegration after the alignment which may pose problems, whenever words' boundaries are placed across a silence. When this happens the reintegration procedure tries to minimise such problems by adjusting boundaries that are close to long silences (this situation usually happens when there are two consecutive words that end and begin with similar sounds and there is a long pause between those words). If silences are short and the boundary is far enough (this situation may happen with very long plosives) the silence information is ignored and the silence duration is considered as part of the word. As reported in Table 1 this strategy always provided the worst results. Despite of this it should be considered that this system made no use of the silence fillers in the orthographic transcription and that it seemed to work better around long pauses.

Results in Table 1 show that the best (and thus, we infer the most reliable) system is still the baseline SONIC aligner. However we noticed that the three systems were making different kinds of mistakes, so we tried to get advantage of all of them implementing a voting policy.

Voting Procedure. As shown in Fig. 2, we represented each word segment as a point specified by its start-time and end-time markers. For each word we evaluated the distances of the word's segments proposed by the three systems and we identified the two closest segments. If the distance was below 200ms the voting procedure chose the mean of the two segments, otherwise it chose the segment of the most reliable system. The system reliability has been assessed on the "training set", using the OM acoustic model (OM_{train}). Results are reported in Table 1.

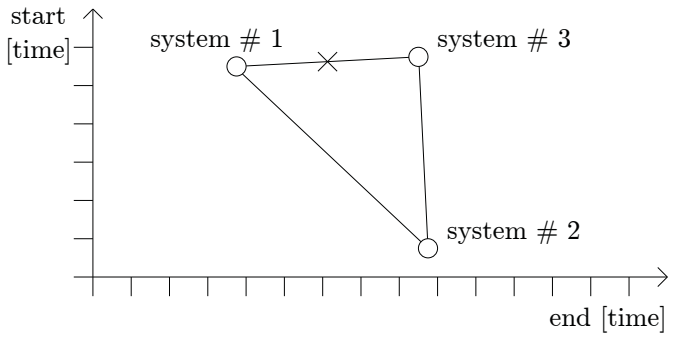

Fig. 2. Word alignment voting procedure

We tested this voting procedure against our development test set, and we saw that it allowed us to gain an absolute 0.2% of alignment correctness, with respect to the most reliable system alone (baseline SONIC). The gain was little but similar across all the measurable configurations (OM and CM acoustic models on both training and development set), even when the actual reliability order did not match the used one (e.g., the $\mathrm{CM}_{\mathrm{AM\ std}}$ in Table 1).

4.5 Phone Alignment

In this task we faced the issue given by words with problematic phonetization in the training set. For example we found the word "`macchina`" with phonetization "`* k *`", with "`*`" being one of the "garbage" fillers in the phonetic vocabulary. Moreover, in the corresponding audio file, intelligible phones could be heard.

For these reasons, and unlike the word forced alignment task, an evaluation process was very difficult to set up, because we couldn't find a reliable reference transcription.

So we used the baseline SONIC (the most reliable system for the word forced alignment task) for this task without any modifications. The output of the system was post-processed in order to comply with the task rules: the vowels were merged together and stress information was discharged.

5 Conclusions

The data provided for the Evalita 2011 word forced alignment task allowed us to train a system and to evaluate its performance. Experiments showed that direct use of external Energy-based SAD in the decoding phase may result in degraded performance, however the use of Energy-based SAD during the training phase significantly improved acoustic model performance for the word forced alignment. It has also been shown that the simultaneous use of several different systems with a proper voting strategy may also improve results. A voting scheme has been proposed that is easy to setup and stable enough to be used successfully.

The Evalita 2011 phone forced alignment task included a phone recognition subtask. Incomplete phonetic transcriptions in the provided data allowed only for suboptimal training and evaluation, nevertheless it was possible to setup a system with reasonable performance.

Acknowledgements. This work has been funded by the EU FP7 ALIZ-E project (grant number 248116).

References

1. APASCI, http://www.elda.org/catalogue/en/speech/S0039.html
2. Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., Omologo, M.: A baseline of a speaker independent continuous speech recognizer of italian. In: EUROSPEECH. ISCA (1993)
3. Black, A.W., Lenzo, K., Pagel, V.: Issues in building general letter to sound rules. In: ESCA Workshop on Speech Synthesis, pp. 77–80 (1998)
4. Black, A., Campbell, N.: Optimising selection of units from speech databases for concatenative synthesis. In: EUROSPEECH, pp. 581–584. International Speech Communication Association (September 1995)
5. Cosi, P., Pellom, B.L.: Italian children's speech recognition for advanced interactive literacy tutors. In: INTERSPEECH, pp. 2201–2204. ISCA (2005)
6. Cosi, P., Tesser, F., Gretter, R., Avesani, C., Macon, M.W.: Festival speaks italian! In: 7th European Conference on Speech Communication and Technology (2001)
7. Gravier, G., Betser, M.: Audioseg (January 2010), https://gforge.inria.fr/frs/download.php/25187/audioseg-1.2.pdf, release 1.2
8. Masuko, T., Tokuda, K., Kobayashi, T., Imai, S.: Speech synthesis using HMMs with dynamic features. In: IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 1, pp. 389–392. IEEE (1996), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=541114
9. Pellom, B.L., Hacioglu, K.: SONIC: The university of colorado continuous speech recognizer TR-CSLR-2001-01. Tech. rep., University of Colorado, Boulder, Colorado (March 2001)
10. Savy, R., Cutugno, F.: CLIPS. diatopic, diamesic and diaphasic variations in spoken italian. In: On-line Proceedings of 5th Corpus Linguistics Conference (2009), http://ucrel.lancs.ac.uk/publications/cl2009/213_FullPaper.doc
11. Yapanel, U.H., Hansen, J.H.L.: A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition. Speech Commun. 50(2), 142–152 (2008), http://dx.doi.org/10.1016/j.specom.2007.07.006
12. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication 51(11), 1039–1064 (2009)