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. 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

Two Forced Alignment (FA) tasks have been proposed by the Evalita 2011 evaluation campaign: the Word Forced Alignment (WFA) and the Phone Forced Alignment (PFA). Two different modalities have been allowed for the tasks, open (OM) and closed (CM).

Three similar systems were tested for the WFA task. The proposed system is based on a voting procedure among them and scored an overall accuracy of 97.90% (OM) and 98.60% (CM). For the PFA task we simply used the best of the three WFA systems which scored 91.30% (OM) and 92.70% (CM).

For the OM we used a previously developed Acoustic Model (AM), that represents our baseline, while for the CM systems we used an AM trained solely with the Evalita 2011 training set. We tested the same training procedure and a slightly different one involving the direct use of Silence/Activity Detection (SAD) algorithm. A description of both the procedures and a comparison is provided within this report.

Our systems have been implemented using Sonic [6], The University of Colorado large vocabulary continuous speech recognition system, and AudioSeg [5], the INRIA audio segmentation and classification toolkit.

2 Data Preparation

One of the main goal of the data preparation step was to obtain a phonetic lexicon, customized for Evalita 2011 data 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" phone belonging to both words. In these cases we converted the shared phone into two distinct phones, each one belonging to a different word, taking as dividing time instant the stop/start word marker.

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 phone that is not related to that letter in the .PHN transcription, and that phone is related with the first (last) letter of the next (previous) word in the transcript, we systematically deleted that phone from the word phonetic transcription.

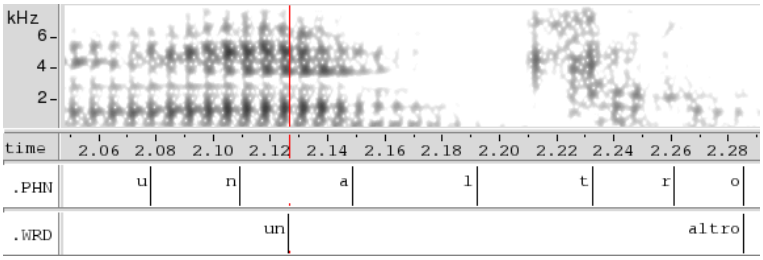

Fig. 1. Example of phones' mismatch due to alignment: the word "un" has been incorrectly phonetized "u n a".

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 (`*` or `#`, for ex.), and the phonetization was made up of several non-garbage phones, out of which intelligible words could be recognized by manual corrections. We set up a procedure for retrieving all those garbage-words whose phonetizations consisted of more than three phones. Those words and related orthographic transcriptions have been manually corrected.

Finally we listed all the words with non alphabetic characters we found in Evalita 2011 train 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>").

2.1 Development Set

We needed a development test environment for internally evaluate our work, so we prepared a "development test" sub set from the Evalita 2011 training data. The corpus have been divided by speaker and sorted by increasing size. Starting from the top of this list, we choose 4 speakers, two male and two female. We decided to have at least one northern and one southern Italian accent speaker.

3 Forced Alignment Procedures

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

3.1 Silence/Activity Detection

In some of our experiments we used an energy based SAD to filter out long pauses that may confuse the automatic systems. Moreover in the Evalita 2011 data, those long pauses usually contains low energy non-transcribed speech that may be even more confusing. The algorithm employed operates in two steps. In the first step it estimates a bi-Gaussian model of the audio frames log-energy, in order to estimate an optimal threshold. In the second step, frames are classified into Silence or Activity accordingly. In our experiments we used a frame length of 200ms, we marked as Activity silence sequences shorter than 100ms and we added a margin of 50ms around all the Activity sequences.

3.2 Acoustic Model

The AM 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 English AM. To make this possible, a phone mapping between Italian and English has been provided, as described in [3]. After alignment, the models are estimated using decision tree state clustering and the procedure is repeated to obtain improved alignments and model parameter estimates. Our AMs consist of gender-independent triphones using standard 39-dimensional PMVDR features.

Open Modality AM In the Open Modality (OM) any type of data can be used for system training, including the provided training set. In our case, the usage of the provided training material was limited to the determination of the reliability of our systems and not for training the AM, that was built upon the APASCI corpus only [1].

Closed Modality AM In the Closed Modality (CM), after the data preparation step, we proceeded with Sonic train procedure, using only Evalita 2011 train data (orthographic transcriptions and audio files) and a phonetic dictionary, that has been built as described in Section 2. We didn't use the provided phonetic information in .PHN files because of missing phones in the transcriptions.

We trained two different AMs. The first system has been built up by audio files that have been processed by the energy based SAD described in 3.1. We avoided the use of timing information of .WRD files in order to demonstrate the validity of the procedure when only the orthographic transcription is available. The second AM was trained without SAD information. We tested the two systems in a WFA 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 (WFA).

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**

3.3 Phonetization Module

Neither the Word Forced Alignment (WFA) nor the Phone Forced Alignment (PFA) 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 hypothesized it. This is especially true for the PFA task, where the task depends on a phone recognition sub-task.

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 [6,2] for missing words. The decision tree was trained with an Italian phonetic lexicon of about 500k Italian forms, originally developed for speech synthesis [4] and then adapted for speech recognition: common alternative transcriptions of some words where added, gemination and syllable division information has been discharged. For the CM the stress marks have been discharged as well.

3.4 Word Alignment

The system used for the WFA task is based on a voting procedure, described at the end of this section, among the following three sub-systems.

$SONIC_{base}$**:** 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 phones 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 3.2, 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 minimize 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 identify the two closest segments. If the distance is below 200ms the voting procedure chooses the mean of the two segments, otherwise it chooses the segment of the most reliable system. The system reliability has been assessed on the "training set", using the OM AM (OM_{train}). Results are reported in Table 1.

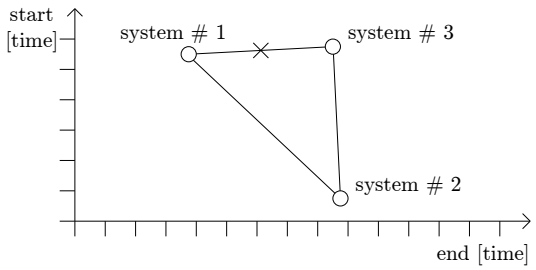

Fig. 2. Word alignment voting procedure.

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 AMs on both training and development set), even when the actual reliability order did not match the used one (e.g., the $CM_{AM\ std}$ in Table 1).

3.5 Phone Alignment

In this task we faced the issue given by words with problematic phonetization in the train 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 WFA 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 WFA) 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.

4 Conclusions

The data provided for the Evalita 2011 WFA task allowed us to train a system and evaluating its performance. We demonstrated that the use of SAD during the training phase significantly improve AM performance for the WFA. 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 PFA task included a phone recognition subtask. Incomplete phonetic transcriptions in provided data allowed only for suboptimal training and evaluation, nevertheless it was enough to setup a system with reasonable performance.

References

1. APASCI, `http://www.elda.org/catalogue/en/speech/S0039.html`
2. 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)
3. Cosi, P., Pellom, B.L.: Italian children's speech recognition for advanced interactive literacy tutors. In: INTERSPEECH. pp. 2201–2204. ISCA (2005)
4. Cosi, P., Tesser, F., Gretter, R., Avesani, C., Macon, M.W.: Festival speaks italian! In: Seventh European Conference on Speech Communication and Technology (2001)
5. Gravier, G., Betser, M.: Audioseg (January 2010), `https://gforge.inria.fr/frs/download.php/25187/audioseg-1.2.pdf`, release 1.2
6. 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)