The SPPAS participation to Evalita 2011

Brigitte Bigi

Laboratoire Parole et Langage, CNRS & Aix-Marseille Universités,
5 avenue Pasteur, BP80975, 13604 Aix-en-Provence France
brigitte.bigi@lpl-aix.fr

Abstract. *SPPAS* - SPeech Phonetization Alignment and Syllabification, is a new tool to automatically produce annotations which includes utterance, word, syllabic and phonemic segmentations from a recorded speech sound and its transcription. This paper describes *SPPAS* algorithms for phonetization and alignment, and evaluations related to the "Forced Alignment on Spontaneous Speech" task of Evalita 2011.

Keywords: speech, phonetization, alignment, annotation, dictionary, hmm

1 Introduction

SPPAS - SPeech Phonetization Alignment and Syllabification, is developped at LPL (Laboratoire Parole et Langage) [1]. It is currently implemented for French, English, Italian and Chinese and there is a very simple procedure to add other languages. An important point for software which is intended to be widely distributed is its licensing conditions. *SPPAS* uses only resources, tools and scripts which can be distributed under the terms of the GPL license.

SPPAS is based on a dictionary look-ups approach for the phonetization and the use of the grammar-based Julius engine for alignment. [2]. A grammar contains sets of predefined combinations of words and contains one or more representations of the distinct phones that make up each word.

EVALITA is an initiative devoted to the evaluation of Natural Language Processing and Speech tools for Italian.In Evalita 2011 the "Forced Alignment on Spontaneous Speech" task was added. Training data is about 15 map task dialogues recorded by couples of speakers exhibiting a wide variety of Italian variants. During the "Forced Alignment" test, participants were asked to produce the alignment from a word transcription level of a 89 utterances data set. This working note is related to the *SPPAS* participation to the Evalita 2011 campaign for the "Forced Alignment" task. Systems were required to align audio sequences of spoken dialogues to the provided relative transcriptions. We participated to both subtasks: phone segmentation and word segmentation. Two modalities were allowed. We participated to the first one only, named "closed" where only distributed data are allowed for training and tuning the system.

The "Forced Alignment"' task included both phonetization and alignment tasks. Phonetization is the process of representing sounds by phonetic signs.

Alignment is the process of aligning speech with these sounds. *SPPAS* is described in section 2. Section 3 is related to the resources we created for the *SPPAS* participation to Evalita 2011: a dictionary and an acoustic model. The evaluation and discussion is presented in section 4. Final results report a correct phoneme alignment rate of 88.4%, and a correct word alignment rate of 96.7%.

2 SPPAS description

2.1 Phonetization

Clearly, there are different ways to pronounce the same utterance. Different speakers have different accents and tend to speak at different rates. The *SPPAS* phonetization process is based on a dictionary solution which consists of storing a maximum of phonological knowledge in a lexicon. Phonetic variants are proposed to the aligner during the alignment stage. The phonetization is then the equivalent of a sequence of dictionary-look-ups. Then, an important step is to build the pronunciation dictionary, where each word in the vocabulary is expanded into its constituent phones. Actually, some words can correspond to several entries in the dictionary with various pronounciations. To deal with such a case, *SPPAS* determines their correct pronounciation during the alignment step because the pronunciation generally can be observed in the speech. The dictionary contains a set of possible pronouciations of words, including accents as *perchè* pronounced as /b e r k e/, and reduction phenomena as /p e k/.

2.2 Alignment

The alignment problem consists in a time-matching between a given speech utterance along with a phonetic representation of the utterance. The goal is to generate an alignment between the speech signal and its phonetic representation. *SPPAS* is based on the Julius Speech Recognition Engine (SRE). Julius was designed for dictation applications, however the Julius distribution only includes Japanese Acoustic Models. But since it uses Acoustic Models trained using the HTK toolkit, it can also use Acoustic Models trained in other languages. To perform alignment a finite state grammar that describes sentence patterns to be recognized and an acoustic model are needed. A grammar essentially defines constraints on what the SRE can expect as input. It is a list of words that the SRE listens for. Each word has a set of associated list of phonemes, extracted from the dictionary. When given a speech input, Julius searches for the most likely word sequence under constraint of the given grammar. The alignment task is a 2-steps process: the first one choose the phonetization and the second one perform the segmentation. Speech alignment requires also an acoustic model in order to align speech. An acoustic model is a file that contains statistical representations of each of the distinct sounds of one language. Each phoneme is represented by one of these statistical representations. Acoustic models were trained with HTK (Hidden Markov Toolkit) by taking the training corpus of speech, previously segmented in utterrances and phonetized.

3 Resources

3.1 The corpus: train/dev segmentation

The train corpus is made of about 15 map-task dialogues recorded by couples of speakers exhibiting a wide variety of Italian variants. Dialogues length ranges from 7/8 minutes to 15/20 minutes. It contains 8063 utterances with word segmentation and phonetic segmentation.

The development corpus was automatically extracted from these data and was not used to train the acoustic model. Extracted files are whose corresponding to the following criteria: last 2 utterances of each speaker in each dialogue and all utterances containing from 100 to 106 phonemes. This development corpus was made of 200 utterances; its duration is 12 minutes 04 seconds. It contains 2373 words, 6282 phonemes, including 689 "_" and 246 "#" which represents 14.88% (it is about the same rate in the train corpus).

Compared to ideal conditions (i.e. like readed words recorded in an anechoic room) we think this corpus added 3 main difficulties: 1/ noisy conditions (the interlocutor recorded in the speaker signal, people walking...) 2/ spontaneous speech (reduction phenomena, hesitations, ...) and 3/ regional accents. During the development stage it was then difficult to analyse which of these phenomena caused errors, and then how to improve algorithms. Consequently, we did not specifically adapted our system.

3.2 Phonetic resources for phonetization

SPPAS made use of the phoneme set proposed in the dialogues phonetization, except for the "#" symbol which we renamed "gb". We also added a specific phoneme "fp" to represent filled pauses for words like "<eh>", "<ah>", etc.

The Italian dictionary was downloaded from the Festival synthetizer tool.The phoneset was automatically modified to match with our Evalita phoneset. This dictionary was enriched by word pronounciations observed in the Evalita train corpus (excluding the extracted development corpus). We corrected manually a large set of these both phonetizations. Finally, the dictionary is made of about 390k words and 5k variants.

3.3 Acoustic Model Training

SPPAS is based on a common practice and uses context-independent Hidden Markov models (HMMs) for speech segmentation. The phoneme statistical representation is based on a 5-states model with a left-to-right topology with self-loops and no transitions which skip over states, as represented in Figure 1 with its initial probabilities. We used context-dependent models such as triphones.

The HMM states are modeled by Gaussian mixture densities whose parameters are estimated using an expectation maximization procedure. The outcome of this training procedure is dependent on the availability of accurately annotated

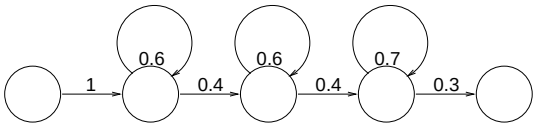

Fig. 1. 5-states HMM with initial probabilities

data and on good initialization. As more speech audio data is collected, better Acoustic Models can be created. We used as input the proposed phonetized transcription, without using the phonetic time-alignment during the training procedure because it is a more generic approach (easier to re-use for other languages). Acoustic models were trained from 16 bits, 16000 hz wav files. The Mel-frequency cepstrum coefficients (MFCC) along with their first and second derivatives were extracted from the speech in the standard way: MFCC_D_N_Z_0.

The acoustic model training procedure was based on 3 main steps. Step 1 is the data preparation. It established the list of phonemes, plus silence and short pauses. It converted the input data (phonetization of the corpus) into the HTK-specific data format. It coded the (Audio) Data: this step is called "parameterizing the raw speech waveforms into sequences of feature vectors" (from wav format to MFCC). Step 2 is the monophones generation. It created a Flat Start Monophones model by defining a prototype model and copying this model for each phoneme. Then, this flat model was re-estimated using the MFCC files to create a new model. Then, it fixed the "sp" model from the "sil" model by extracting only 3 states of the initial 5-states model. Then, this model is re-estimated using the MFCC files and the phonetization. Step 3 created tied-state triphones from monophones and from some language specificities defined by the way of a configuration file. This file summarizes Italian phonemic informations as for example the list of vowels, liquids, fricatives, nasals or stop. We created manually this resource (tree.hed).

4 Evaluation

The most common and direct form of evaluation is comparing the automatic segmentation to a manual segmentation. This evaluation is performed using the proposed phonetizations and alignments (which are supposed to be the right ones) on the development corpus we extracted. The quality of the generated annotation depends largely on the robustness of the HMM recognizer, and on the dictionary. The evaluation we propose is made of 3 stages. First of all we evaluated the alignment only, from the given phonetization. Next, we evaluated the phonetization only. Finally, we evaluated the whole process made of phonetization plus alignment. These evaluations was performed using tools we developped or using Sclite [3]. Accuracy was calculated as a function of words or phonemes.

Acoustic model evaluation:
In this experiment, the automatic alignment was estimated on the basis of the manual phonetization. Table 1 proposes detailed alignment performances depending on the delta range between the manual and the automatic alignment, by using the time-location of the middle of each phoneme. The system produced 88.4% of correct alignments with a delta range of 60ms and only 3.6% outliers.

Table 1. Alignment-only *SPPAS* performances

Delta	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.20
Nb Corr	2513	3929	4712	5160	5406	5555	5680	5753	5804	5859	6055
% Corr	40.02	62.55	75.09	82.14	86.05	88.43	90.42	91.58	92.39	93.26	96.39

Evaluation was also performed with Sclite using the time-alignment option. It reports a correct rate of **89.8%**, with 7.6% substitutions, 2.6% deletions and 2.6% insertions.

Figure 2 is a boxplot of the most frequent phones. It represents the delta between the phone durations of the automatic alignments and the phone durations of the manual alignments. It shows that in general the automatic system produced vowels with a smaller duration. It also shows that pauses, filled pauses and garbage have the greatest ranges. This is probably due to the recording conditions (noises are frequents) and to some aspects of spontaneous speech.

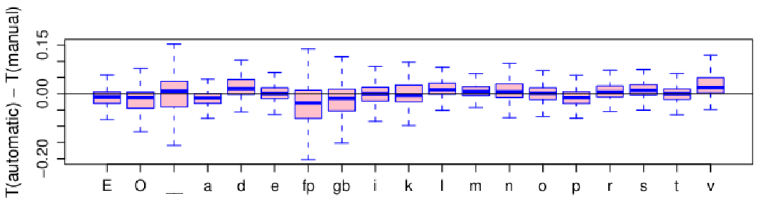

Fig. 2. Phones delta durations for the most frequent phones

Phonetization evaluation:
We propose an evaluation of the phonetization only, which represents the availability of our system to propose the "good" phoneme sequence. This evaluation was estimated with Sclite but without any time-alignment constraints. It reports a correct rate of **89.5%**, with 8.1% substitutions, 2.3% deletions and 6.9% insertions. Most frequent errors are due to the garbage manual annotation: in this case the system anyway proposes a phonetization. For example, the sentence "bravissimo a questo" is phonetized as "b r A V I s I M O A k we s t o" but

the manual phonetization is “b r # s # k we s t o” which generates 5 insertion errors and 2 substitutions. The same is for the phonetization “d I R E T A M E n t e” as the manual one is “d # D # n t e” that generates 4 insertion errors and 3 substitutions. Many other errors are due to phenomena frequently observed in spontaneous speech. One of the characteristics of Spontaneous Speech is an important gap between a word’s phonological form and its phonetic realisations. Specific realisation due to elision or reduction processes (for example *perchè* pronounced as /b e k/, *il videotelefono* as /jo d e l e f/) are frequent in spontaneous data. It also presents other types of phenomena such as non-standard elisions, substitutions or addition of phonemes which intervene in the automatic segmentation. A set of these instances can be added to the dictionary but it will not cover all the possible observed realisations. We think this is the main limitation of the dictionary-based phonetization approach.

5 Conclusion

This working note presented *SPPAS*, a tool to perform the forced-alignment task during the Evalita 2011 campaign, on Italian map-task dialogues. It described the development of resources and free tools, consisting of acoustic models, phonetic dictionaries, and programs to deal with these data. Figure 3 is a Screenshot of *SPPAS* alignments. It is important to mention that *SPPAS* can deal with various languages and it was not specifically devoted to Italian.

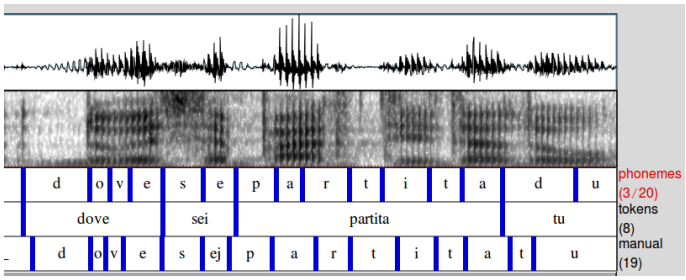

Fig. 3. *SPPAS* output example

References

1. Bigi, B.: SPPAS: SPeech Phonetization Alignment and Syllabification, http://www.lpl-aix.fr/∼bigi/sppas/ (2011)
2. Nagoya Institute of Technology: Open-source large vocabulary csr engine julius, rev. 4.1.5 (2010)
3. Speech Recognition Scoring Toolkit: http://www.itl.nist.gov/iad/mig/tools/, version 2.4.0 (2009)