Semi-automatic phonetic labelling of large corpora

Odile Mella, Dominique Fohr

SEMI-AUTOMATIC PHONETIC LABELLING OF LARGE CORPORA

O. Mella and D. Fohr

CRIN-CNRS & INRIA Lorraine
Batiment LORIA, B.P. 239
F54506 Vandoeuvre-lès-Nancy, France
Tel. +33 3.83.59.20.80 Fax. +33 3.83.41.30.79 E-mail: {mella,fohr}@loria.fr

ABSTRACT

The aim of the present paper is to present a methodology to semi-automatically label large corpora. This methodology is based on three main points: using several concurrent automatic stochastic labellers, decomposing the labelling of the whole corpus into an iterative refining process and building a labelling comparison procedure which takes into account phonologic and acoustic-phonetic rules to evaluate the similarity of the various labelling of one sentence. After having detailed these three points, we describe our HMM-based labelling tool and we describe the application of that methodology to the Swiss French POLYPHON database.

1. INTRODUCTION

Training and assessment of speech recognition systems, especially those based on Hidden Markov Models and Artificial Neural Networks, need the availability of large speech corpora. Furthermore, most of the continuous speech recognition systems used phoneme-like units. Therefore, the corpora have to be reliably phonetically labelled, that is a phonetic transcription and an accurate alignment have to be provided. Two approaches have been mainly used for this purpose hand-labelling and semi-automatic.

Both methods have advantages and drawbacks. Hand-labelling allows both fine phonetic transcription and accurate boundaries. By contrast, this task is tedious, time consuming and may lead to a lack of homogeneity when several labellers are involved. For huge corpora, hand-labelling is not tractable, so automatic labelling is the only practicable solution. Moreover, an automatic procedure achieves consistent alignment. But, the major problem is that gross errors may occur, mainly because of the differences between the actual utterance and the generated phonetic transcription like deletions, liaisons,... For this reason, the results of the automatic labelling require to be manually verified [2].
These observations have led us to elaborate a methodology to label large corpora which speeds up and reduces the step of manual verification.

2. METHODOLOGY

Our methodology tries to solve two main problems in labelling large corpora of several thousand sentences: how to label them with systems using an automatic training stage, like HMM or ANN systems, which usually need large training corpora to be efficient and how to evaluate the result of this labelling.

2.1 Evaluation of the labelling

To evaluate the automatic labelling, our basic idea is to use two or more independent labelling tools [1] and to design an algorithm to compare their results in order to classify them in two categories. When the two sequences of labels and their associated boundaries provided by the two systems match, the corresponding speech signal is assumed as correctly labelled and no further manual correction will be necessary. The remaining sequences are rejected and they will have to be either manually checked and corrected or re-labelled by the automatic labeller, if it can be improved. The comparison algorithm and the matching criteria will be explained in the section 3.

2.2 Iterative labelling process

In order to allow an automatic-training-based system to provide an efficient labelling without large training corpora, we propose an iterative refining process which may be broken down into several stages:

1. In order to train the system, a small part A of the corpus have to be hand-labelled.
2. Another part B bigger than A is automatically labelled by the system.
3. For every item, for instance a sentence, the result of the labelling process is evaluated as correct or rejected by comparing the generated sequence of labels and its alignment with those generated by another labelling tool.
4. The sentences marked as correct are used to retrain the labelling system.
5. (optional step) The rejected sentences can be manually corrected and added to the training corpus.
6. The process is iterated from step 2 with a bigger part of the corpus until it will be fully labelled.

Using an iterative process leads to more accurate labelling systems. For instance, with a HMM-based system, during the iterative process, the number of probability density functions (pdf) can be increased or new models such as context-dependent phoneme models can be trained. Moreover, the study of the mislabelled sentences allows to improve and fix some parts of the labelling tool.

3. COMPARISON ALGORITHM

The comparison algorithm is the main part of our methodology. Its basic aim is, considering a sentence, to determine the similarity of both results of its labelling provided by two different labelling tools. Such a procedure must not depend on the two labelling tools (lexicon, transcription rules, accuracy of labelling,...). and must allow every user to specify his criteria of similarity. Thus, it must be customisable.
The procedure is composed of three steps: a rewriting step, an alignment algorithm and a procedure of decision making.

3.1 Conversion of the input strings

As the two labelling tools may use nonuniform sets of phonetic symbols, the user can define a common phonetic alphabet and the corresponding rewriting rules.

3.2 Alignment algorithm

To compare the two labelling results of a sentence, we firstly need to perform an alignment of both strings of labels, which can have unequal lengths. So, we have designed a classic elastic comparison algorithm (DTW), but, in order to help the alignment process, the user can define:

- the available insertion cases; in other words, a list of phonetic symbols corresponding to sounds which can be frequently deleted or inserted in the utterance,
- the substitution cases which do not result in penalties, as a list of couples of phonemes that the user judges similar.

The two first lines of Table 1 displays the result provided by the alignment algorithm, that is the paired labels and the inserted (or deleted) labels if there are.

3.3 Decision making procedure

3.3.1 General principle

After aligning the two labelling results of a sentence, the comparison algorithm must determine if both sequences of labels are close or not. For this purpose, the decision making procedure browses the alignment result, and, compares every couple of labels pairing by the DTW algorithm, if it is necessary, takes into account the inserted labels, and uses comparison rules in order to generate equivalent groups of labels. In the example of Table 1, the groups of labels / i e / and / i j e / are classified as equivalent. We will explain in the paragraph 3.3.2 how they become equivalent.

Then, the procedure checks the shifts of the extreme boundaries of every equivalent groups of labels, to determine if both groups definitively match. At this end, the user can define for each label the allowed maximum shifts of the beginning and end boundaries.

In every sentence, groups of labels are marked as mislabelled or as well-labelled and finally the sentence is rejected or not.

3.3.2 Comparisons rules

To be general, our comparison tool needs to know the degree of similarity between the results provided by both labellers wished by the user. Purposely, the comparison of labels or of groups of labels operates with ordered phonological and acoustic-phonetic rules given by the user. These rules specify the available differences between two different groups of labels.

Mainly, these differences can be due to:

- the distinct sets of phonemes used by the labellers; for instance, in French, three or four nasal vowels can be used;
- the differences between the lexicons on which the labellers are based;
- the multifarious rules and procedures applied by the labellers for the generation of all the potential phonetic realisations from the same orthographic transcription; namely, how assimilations, deletions, insertions, liaisons, allophones, infra-phonemic segments and extra speech segments are taken into account;
- the various aims of labelling: to label the sounds actually uttered or what the speaker has intended to pronounce. For example, in French, the standard transcription of the word «*médecin* » is /mɛdse~/ but it can be also pronounced / mɛtse~/ or /mɛd@se~/;

3.3.3 Examples of rules

These phonological and acoustic- phonetic rules can be categorised as in the following list. It should be noted that the rules are formulated like rewriting rules but the groups of labels are indeed not rewritten, they are only compared. As follows, we show some of the implemented rules. To make their understanding easier, we prefer particularising a rule with examples of phonemes even if the rule is available for a class of phonemes.

- *archiphonemes*

Such rules are needed when the phonetic alphabets or the lexicons are distinct or when the labelling tools do not have the same accuracy:

[e ⇒ ai] ; [ai ⇒ e] ; [e ⇒ E] ; [ai ⇒ E]

- *deletion of French schwa*

The schwa deletion very often happens in French. Thus, lexicons may generate several potential utterances for one sentence. According to the schwa duration or to the accuracy of their models the labelling tools may obtain

two different sequences of labels. Here '*' means any phoneme.

[* @ ⇒ *]

- *allowed insertions*

We could discriminate two types of allowed insertions, those arisen from coarticulation phenomena (α) and those introduced by the accuracy of the labeller (β):

[i j a ⇒ i a] (α)

[**!** * ⇒ *] (β)

- *double phonemes*

The word concatenation may lead two sequential phonemes to be uttered as only one:

[a a ⇒ a] ; [d d ⇒ d]

- *assimilation rules*

Depending on whether the labelling tool labels the sounds actually uttered or what the speaker has intended to pronounce, the comparison needs to deal with assimilation rules such as:

[e~ **t** k] ⇒ [e~ **n** k]

[**t** d ⇒ **d** d]

3.3.4 Boundary checking

As it has been previously introduced, after making the equivalence between two labels or two groups of labels, the comparison algorithm checks the shifts of the extreme boundaries to determine if both groups are definitively equivalent.

To be still general, the procedure requires that the user gives the allowed maximum shifts of the beginning and end boundaries between groups of labels. These limits obviously depend on the phonemes or class of phonemes but they are also context-dependent.

4. LABELLING TOOL

Our labelling tool is based on second order Hidden Markov Models with 35 context-independent phonemes. Each model contains 3 states, left-to-right, no skip, self-loop with initially one probability density function (pdf) per state [3]. The speech parameters are 12 MFCC coefficients plus first and second derivatives using a mean cepstre removal computed on the whole sentence.

From the orthographic transcription of every sentence, from the phonetic lexicon BDLEX and from a set of phonological rules, we generate all the potential phonetic realisations. Our aim is to label what the speaker has intended to pronounce and not exactly the sounds uttered. So, we do not take into account assimilation rules as nasalisation or unvoicing. By contrast, pauses, French schwa deletions or insertions and liaisons are taken into account.

The system performs a forced alignment between the speech signal and all the potential phonetic realisations. The one with the best alignment score is retained as the labelling sequence.

5. LABELLING of POLYPHONE

5.1 POLYPHONE database

We have applied the previously described methodology to label the Swiss French POLYPHONE[1] database. This database is made up of telephone recordings of sentences from 4500 speakers recorded over the SwissNet by the SWISS TELECOM PTT and the IDIAP laboratory. The speech files are in format A-law, 8 bits, 8kHz. The orthographic transcription of the actually uttered sentences is supplied.

We have decided to label all the phonetically rich sentences of this corpus, that is 45000 sentences.

5.2 Application of the labelling methodology

With respect to our methodology we had to hand-label a part A of POLYPHONE database in order to train the automatic labellers. We have chosen to use another French corpus (BREF 80) already well-labelled to train our HMM phoneme models for male and female speaker. Because this corpus was not recorded over the telephone, it has been bandpass filtered (330-3400 Hz).

Likewise, a part B of POLYPHONE database had to be automatically labelled by our labelling tool and by the labeller of another laboratory. As we did not have the results of the other labeller, we have decided to replace it by an hand-labelling. So, we have manually and automatically labelled one hundred POLYPHON sentences.

These hundred sentences have allowed us to test our comparison algorithm and to assess the comparison rules and the maximum shifts for the boundaries.

Figures 1 and 2 display two spectrograms of the speech signal with the results of the hand-labelling (upper alignment) and the automatic labelling (lower alignment). Figure 1 shows a part of a sentence which has been classified as well-labelled by the comparison algorithm. Indeed, the biggest differences between the two labelling results: the shift of the end boundary of / R/, the /j/ insertion (see 3.3.3) and the insertion of schwa could be considered as acceptable. By contrast, the labelling evaluation procedure has rejected the part of a sentence presented in Figure 2. The too big shifts of the boundaries of the phoneme /Z/ underline the misalignment of the phoneme by the automatic tool.

We have tested the feasibility of our automatic labelling checking procedure, assessed the rules and our HMM-based labelling tool on this first part of the Swiss POLYPHONE database. We are now applying our methodology to label the 45000 sentences of the corpus.

[1] POLYPHON database belongs to SWISS TELECOM PTT and we use it according to a convention between CRIN and SWISS TELECOM PTT and IDIAP.

6. CONCLUSION

We have elaborated a methodology to semi-automatically label large corpora of several thousand sentences. This methodology is based on using several concurrent automatic labellers, applying an iterative refining process and using a labelling comparison algorithm in order to classify the sentences into well-labelled and mislabelled.

The parameterisable comparison algorithm is the main part of our methodology. It does not depend on the two labelling tools and allows every user to specify his criteria of similarity by specifying a set of phonological and phonetic rules. By changing these rules, our comparison algorithm can be completely adapted to assess the labelling results in other languages [1].

7. REFERENCES

[1] A.Vorstermans, J.P. Martens and B. Van Coile, « *Fast Automatic Segmentation and Labeling: Results on TIMIT and EUROM0* »,Proceedings of 4th European Conference on Speech Communication and Technology, Madrid 1995.

[2] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, M. Omologo, « *Automatic Segmentation and Labeling of English and Italian Speech Databases* », Proceedings of 3rd European Conference on Speech Communication and Technology, Berlin 1993.

[3] J.F Mari and D. Fohr and J.-C. Junqua, « *A second-Order HMM for High Performance Word and Phoneme-Based Continuous Speech Recognition* », Proceedings of International Conference on Acoustics, Speech and Signal Processing, Atlanta 1996.

Labeller 1	#	i	l	v	a	f	a	l	w	a	R	p	l	i		e	b	a	g	a	Z		#
Labeller 2	#	i	l	v	a	f	a	l	w	a	R	p	l	i	j	e	b	a	g	a	Z	@	#
Boundary shift in ms	17	18	2	1	3	3	4	9	54	3	71	1	12			8	1	6	18	6	42		

Table 1. Example of alignment between the results from two different labellers.

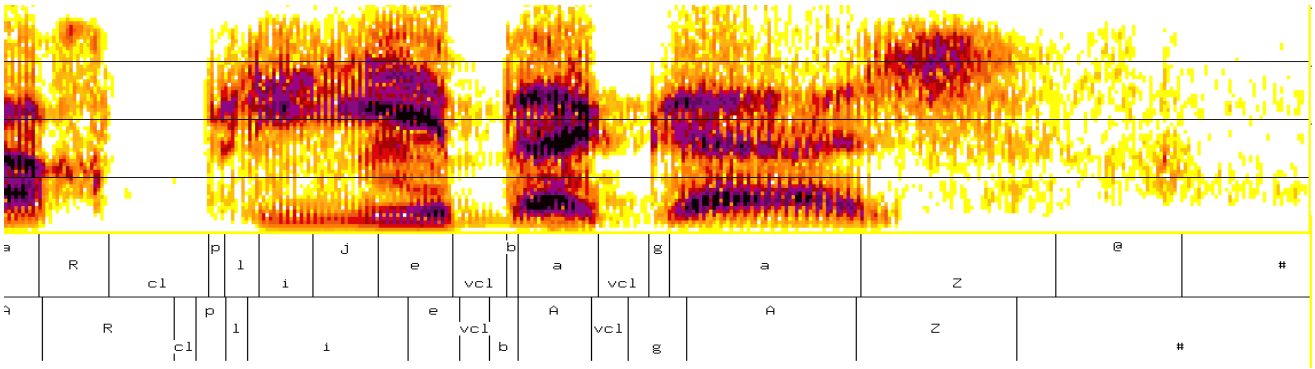

Figure 1. The two labelling results associated with the alignment displays in Table 1

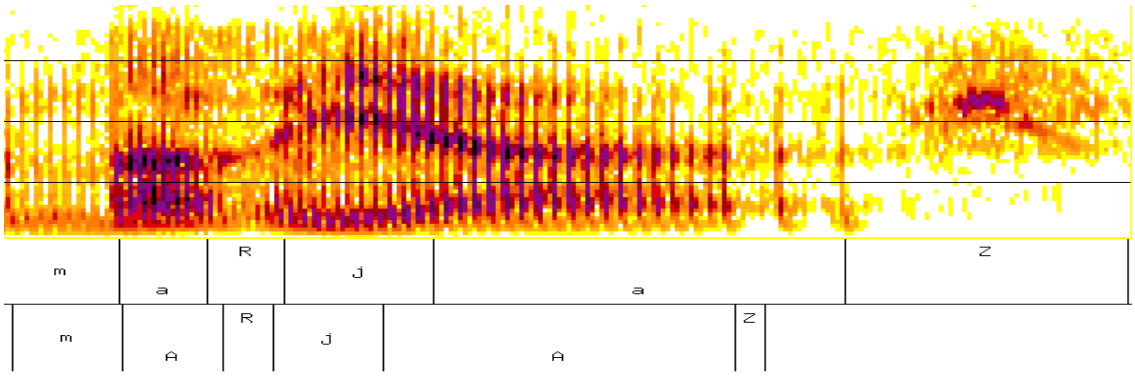

Figure 2. Another example of comparison of labelling results.