Two Tools for Semi-automatic Phonetic Labelling of Large Corpora

Odile Mella, Dominique Fohr

Two Tools for Semi-automatic Phonetic Labelling of Large Corpora

Odile Mella and Dominique Fohr

LORIA-CNRS & INRIA Lorraine & UHP
BP 239 F54506 Vandoeuvre France
mella,fohr@loria.fr

Abstract

This paper presents two tools allowing a reliable semi-automatic labelling of large corpora : an automatic HMM-based labelling tool and an assessment and decision system to validate the automatically labelled sentences. This decision system uses the results supplied by another automatic labeller and compares their results with a parametrisable comparison process. We also propose an generic methodology to improve the labelling accuracy and to reduce the step of manual verification.

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 use phoneme-like units. Therefore, the corpora have to be reliably phonetically labelled, that is a phonetic transcription and a accurate alignment of this transcription on the speech signal have to be provided. Two approaches have been mainly used for this purpose:

- hand-labelling, realizing simultaneously the phonetic transcription and the time alignment,
- semi-automatic labelling, which is composed mainly in three steps: providing a phonetic transcription from an orthographic string, then aligning this sequence of phone labels with the speech signal.

Hand-labelling allows both a fine phonetic transcription and accurate boundaries. But, this task is 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 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 (Vorstermans & Martens & Van Coile, 1996; Depambour & al., 1997).

The purpose of this paper is triple: to present an automatic labelling tool and to describe a generic process to label semi-automatically large corpora and two methods to speed up and to reduce the step of manual verification.

2. Labelling tool

Given the speech signal and the orthographic transcription of a sentence, this labelling tool (labeller) provides a sequence of phonetic labels with associated begin-end boundaries. It is composed of two main parts: a generator of potential phonetic transcriptions of a sentence and a alignment program of these transcriptions on the speech signal.

2.1. Phonetic transcription generator

2.1.1. Introduction

The purpose of a phonetic transcription generator is to provide a phonetic transcription from the orthographic transcription of a sentence or a text. But, a sentence can be uttered in several phonetic realizations. Let us quote some French examples. First, a orthographic transcription can have more than one phonological transcription: for instance, the word *"jean"* must be pronounced /dʒin/, if it means a item of clothing and /ʒã/ if it is the French first name. Secondly, a speaker can or must insert a phoneme of liaison between two words: the definite article *"les"* is uttered /le/ when it is followed by a consonant and /lez/ when it is followed by a vowel. Furthermore, according to speaking rate, accents and dialects, some phonemes can be omitted, like the French schwa: the adjective *"petit"* can be pronounced /pəti/ or /pti/. Finally, coarticulation phenomena result in alterations of phonemes like voicing or nasalisation: *"Banque de France"* can be pronounced /bãkədəfrãsə/ or /bãŋdəfrãs/.

Therefore, as the actual utterance of a sentence by a speaker is unknown, the generator must be able to provide a great number of potential phonetic realizations from the orthographic transcription of a text, or at least the usual ones.

2.1.2. Principle

Our phonetic transcription generator has for input a ASCII file containing the orthographic transcription of a sentence and products a phonetic graph giving several phonetic transcriptions of this sentence as shown in Figure 1.

For that purpose, our generator uses the French BDLEX lexicon developed by IRIT completed by an application-specific lexicon and carries out the following tasks:

- it translates numbers, units and currencies in full, like the string *"22F"* into *"twenty-two francs"*;
- for every word in the sentence, the generator extracts all of its phonetic transcriptions from the lexicons in two passes : one respecting the case, the other translating the word in lower case. Moreover, if a compound (i.e. containing an hyphen) is not found, the search is retried with each of its component words;
- the system combines the several phonetic transcriptions of all the words of the sentence into a graph, taking into account :
  - the multiple realizations of a word,
  - the possible liaisons between words,
  - the optional deletion of the French schwa,
  - the optional insertion of a pause between words.

The building of this graph is explained in the next paragraph.

2.1.3. Phonetic graph building

In French, the liaison between two words happens if the second word begins with a vowel. However, this liaison is not always mandatory. Thus, our system only allows two transitions between the two words: one including the liaison consonant the other one including a pause between both words ; that is to say, either the speaker uttered the liaison or he inserted a pause.

In addition, the insertion of a liaison consonant usually do not change the realization of the previous vowel except in few words ending by a nasal vowel. For instance, the word *"bon"* /bõ/ becomes /bɔn/ when it is followed by a word beginning by a vowel like in the phrase *"bon ami"*. Our generator copes with these exceptions.

With regard to deletions, our generator is able to take into account the deletions coded in the lexicons. But the lexicon that we have used only coded the optional deletion of the French schwa at the end of a word. Therefore, we have added a specific module to deal with the deletion of the schwa in the adverbs ending by *"ement"* which often occurs in French.

As we cannot predict when the speaker pauses for breath, we have chosen to put an optional pause after every word.

Figure 1 shows the phonetic graph generated from the orthographic transcription *"Mon ami Jean lit rapidement"*. It can be noticed:

- the double transition between *"mon"* and *"ami"* with the liaison consonant /n/ or a pause /#/,
- both potential phonetic transcriptions of the French word *"jean"*,
- the optional pauses between words,
- the possible deletion of the /ə/ in the adverb *"rapidement"*.

2.1.4. Conclusion

To summarise, the transcription generator, the first part of our labelling tool provides a set of potential phonetic transcriptions. Moreover, its aim is to label what the speaker has intended to pronounce and not exactly the sounds uttered. Thus, it takes into account optional pauses, liaisons and French schwa deletion but it does not take into account assimilation phenomena as nasalisation or unvoicing.

2.2. Alignment algorithm

The second part of the labelling tool performs a forced alignment between all the different paths of the phonetic graph and the speech signal. The path obtaining the best alignment score is retained as the labelling result.

The alignment algorithm is based on second order Hidden Markov Models. It uses one HMM per phoneme and one more for the pause. It works with 35 context-independent models because we have chosen not to discriminate certain phonemes like those belonging to a phonological opposition that it can be neutralised, like the nasal vowels /ɛ̃/ and /œ̃/.

Each HMM model is composed of 3 states whose the topology is: left-to-right, no skip, self-loop. One probability density function (pdf) with a full covariance matrix is estimated per state.

The speech parameters are 12 MFCC coefficients plus first and second derivatives using a mean cepstre removal computed on the whole sentence.

The Baum and Welch algorithm is used for the training of the models and the Viterbi's one for the alignment (Mari & Fohr & Junqua, 1996).

Figure 2 presents the alignment path for the sentence *"Mon ami Jean lit rapidement"*, the aligned graph of which is shown on Figure 1. Finally, Figure 3 displays the spectrogram of the sentence and its labelling results (labels and beginning-end boundaries) provided by the labelling tool.

3. Methodologies for semi-automatic labelling of huge corpora

3.1. Introduction

As it has been introduced in section 1, the phonetic labelling of huge corpora needs a automatic labelling tool. But automatic labelling induces two problems. First, the automatic labelling tools are often based on statistical methods needing an automatic training stage, which itself requires a large corpus already labelled. Above all, in automatic labelling, errors and even gross errors may occur and make necessary the step of manual checking of the labelling results.

A part of these errors may result from the differences between the actual utterance of a text and the phonetic transcription generated by the labelling tool (see section). However these errors may also be caused by:

- the occurrence of extra speech (noise, cough, laugh,...),

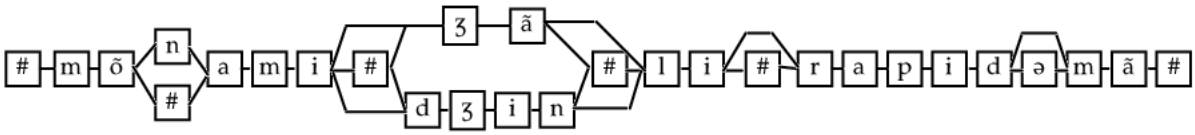

Figure 1: The phonetic graph of the sentence *"Mon ami Jean lit rapidement"*

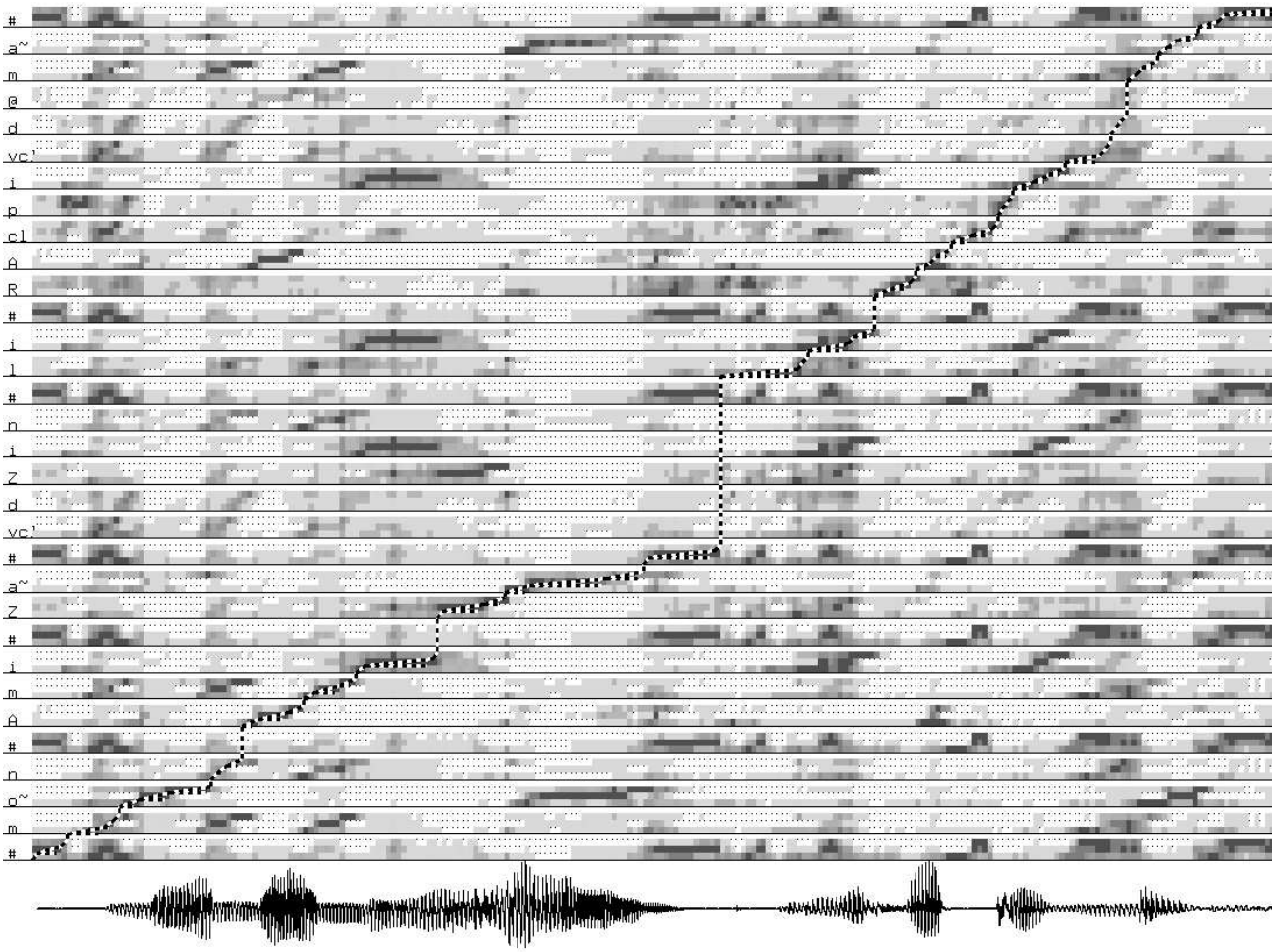

Figure 2: The alignment path between the speech signal and the HMM models corresponding to the phonetic graph shown on the Figure 1.

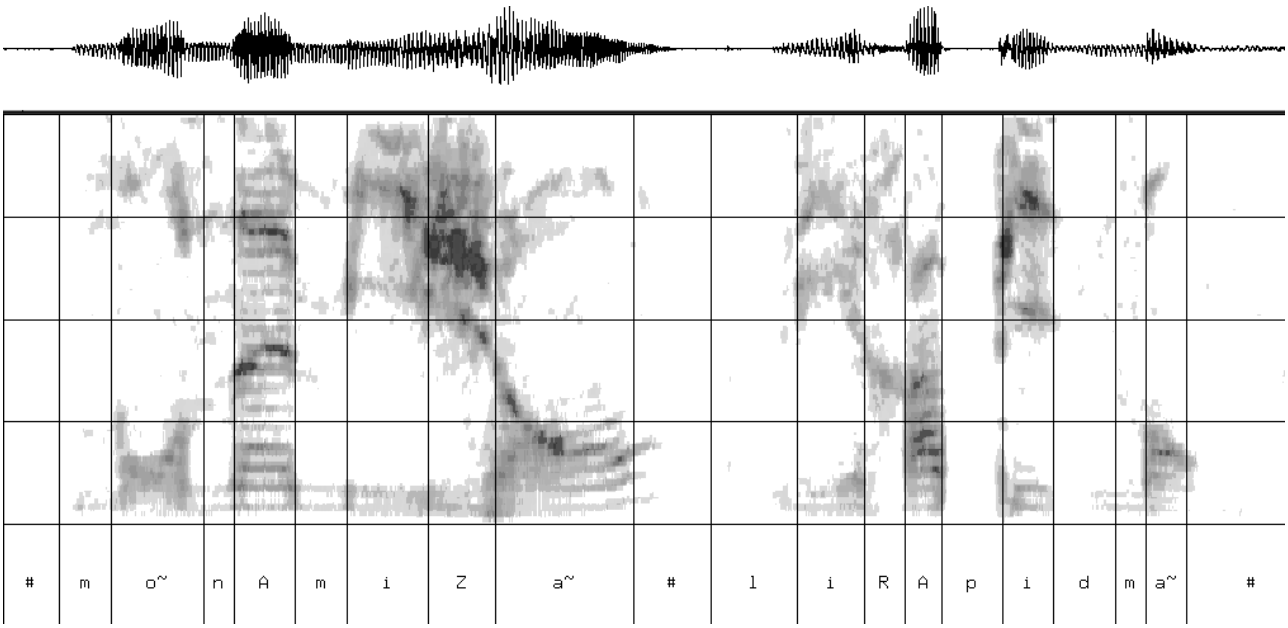

Figure 3: Results of the labelling of the sentence

- the mistakes made by the speaker,
- in the case of training-based labelling tools, the mismatch of the acquisition conditions between the training corpus and the corpus to label (microphone, chanel, speaker, dialect,...).

Thus, in the following sections, we propose, on the one hand, a generic methodology to automatically label large corpora with automatic-training-based systems and, on the other hand, two methods to assess the automatic labelling in order to reduce to the minimum the stage of manual verification of labelling results.

3.2. Generic methodology to label large corpora

We have elaborate an iterative refining process which permits a training-based labelling tool to provide a final efficient labelling without requiring a training of the tool on a large similar corpus (namely with the same acquisition conditions)

This iterative process may be broken into several steps, given a corpus of several thousands of sentences:

1. Training the labeller on a bootstrap corpus which may be:
   - either a small hand-labelled part of the corpus to label,
   - or any labelled corpus else with different conditions of acquisition;
2. Labelling all the corpus or a part A of the corpus;
3. Evaluating the labelling results of every sentence and classifying it as correctly labelled sentence or mislabelled sentence by an assessment method;
4. Using the sentences marked as correct to retrain the labelling tool;
5. Iterating the process from the step 2 with a bigger part of the corpus or the whole corpus until it will be fully correctly labelled or until the amount of mislabelled sentences does not decrease any more.

The remaining mislabelled sentences either will have to be manually checked and corrected or may be labelled again if the system may be improved.

The main part of this refining process is the labelling assessment method. If two different automatic labelling tools are available, we propose to verify the labelling by comparing the results of the two labellers with a customisable comparison process. If, you unfortunately have only one automatic labeller, you could use another verification methodology. It is roughly based on the analysis of the alignment scores provided by the labeller. These two methods will be detailed in the following sections.

3.3. Labelling assessment with two labelling tools

The major idea of this labelling assessment methodology is based on the comparison of the results of the two automatic labellers. A comparison procedure determines, for every sentence, the similarity of both results provided by the labelling tools. In other words, when the two sequence of labels and their related boundaries provided by the two systems match, the corresponding speech signal is deemed as correctly labelled and no further manual correction will be necessary.

This procedure must be as generic as possible and must not depend on the features of the labelling tools : lexicon, transcription rules, phonetic alphabet, requirements of labelling accuracy (phonemic, phonetic,...). For this reason, the comparison process is composed of three steps: a rewriting algorithm, an alignment algorithm and a decision

making procedure.

3.3.1. Rewriting algorithm

As the two labelling tools can use non uniform sets of phonetic symbols, the user can define a common phonetic alphabet and the corresponding rewriting rules. These rules merely build a larger set of phonetic symbols and certainly do not have a comparison role. Here are two examples of rewriting rules:

[ɛ => ai]
[tcl t => t burst]

3.3.2. Alignment algorithm

The alignment algorithm begins the phase of the comparison of labelling results of both automatic labellers. It tries to pair the two sequences of labels considering that they could not have the same length due to deletions or insertions and that the labels could be different because both labellers do not use the same lexicon. Because of this, this alignment algorithm is based on an elastic comparison algorithm (DTW) between the two strings of labels.

In order to guide the alignment process, the user can indicate a set of phonemes or sounds which are often inserted or deleted by the labellers, like the French schwa /ə/, the French /j/ or extra speech symbols. The user provides them with an insertion/deletion matrix.

In the same way, he can give the couples of phonemes which can be paired although they are different; especially phonemes which are acoustically close and belong to a phonological opposition that it can be neutralised, like the vowels /e/ and /ɛ/ in the French word *"maison"*.

As results, the algorithm provides the best alignment path between the two sequences of labels.

3.3.3. Decision making procedure

Afterwards, the decision making procedure determines which parts of the sentence are correctly labelled, namely, for which parts both tools have provided similar labelling results. For that, the procedure backtracks the alignment path, compares every couple of labels paired by the alignment algorithm, takes into account the inserted/deleted labels, and finally generates **equivalent** groups of labels.

Two groups of labels are deemed as equivalent, if:

- either they have the same number of elements and all the elements are identical,
- or their elements are different but the confusions, insertions or deletions which cause the difference are allowed by the user with **comparison rules**.

Finally, the decision making process checks the shifts of the beginning and end boundaries of every couple of equivalent groups of labels to determine if they are similar and thus correctly labelled.

3.3.4. Comparison rules

The comparison rules given by the user have two functions. On the one hand, they make both labelling results comparable, that is they adapt the phonetic accuracy of the most accurate labeller to the less accurate one's. For instance:

- one of the two labellers splits a plosive segment into a closure part and a burst part and the other does not:
  [tcl t => t]
- one of the two labellers discriminates /e/ from /ɛ/ the other does not:
  [e => E]
  [ai => E]

On the other hand, the comparison rules specify the degree of similarity wished by the user, ie. the allowed differences between two equivalent groups of labels. For instance:

- both labellers discriminates /e/ from /ɛ/ but the user does not consider the confusion of these phonemes as a gross error:
  [e => ai]

In both cases, the comparison rules deal with the differences between two labelling results, thus we categorise these rules according to the sources of these differences. In addition, we present some s examples of implemented rules. 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.

The phonetic accuracy of the labelling: the two labellers may not have the same phonetic accuracy, namely the aligned transcription may be a phonological transcription, a broad phonetic or an accurate phonetic transcription with allophones, infra-phonemic segments, and extra speech segments. Here are 3 examples:

- the labellers do not discriminate the same set of phonemes, for instance one discriminates the two nasal vowels /ɛ̃/ and /œ̃/, the other does not:
  [œ̃ => ɛ̃]
- one of the two labellers splits a plosive segment into a closure part and a burst part and the other does not:
  [tcl t => t]
- one of the labellers detects some extra speech segments like noises (* means a joker and ! means noise):
  [* ! => *]

The lexicons: the lexicons on which the labellers are based can code differently some words.

- the French word can be transcribed /apyi/ ou /apYi/ where /Y/ is a short /y/:
  [Y i => y i]

The phonetic transcription rules. Even if both labellers work with the same lexicon and the same phonetic symbols, they may differ by the rules and procedures used in the generation of the potential phonetic realizations. In other words, how far deletion insertion and liaison phenomena are taken into account. For instance:

- one of the labellers accept the deletion of the French schwa, the other does not:

  [* ə => *]

- only one of the labellers copes with the insertion of /j/ in the coarticulation of two words when the first of which ends with /i/ and the second of which starts with a vowel:

  [i j a => i a]

Aims of the labelling: A labelling tool is often designed according to the purpose of this labelling: is it to label the sounds actually uttered or what the speaker has intended to pronounce ? How are assimilation and alteration phenomena taken into account ?

- one of the labellers deals with voicing/unvoiving assimilation, the other does not, like in the expression *"sept de cœur"*:

  [t d => d d]

- one of the labellers deals with double phonemes at word boundary, like in *"il alla à Paris"*:

  [a a => a]
  [d d => d]

- one of the labellers taken into account nasalisation phenomena, as in *"Pentecôte"*:

  [ã t k => ã n k]

Labelling errors: Of course, one of the sources of differences between the results provided by several labellers is the errors made by the labelling tools. If the user deems an error as minor, like the confusions between /ɛ/ and /e/, he adds the corresponding comparison rule. Otherwise, the error is major, as missing a liaison between two words, and in this case, the groups of labels will have to be marked as non-equivalent.

3.3.5. Boundary checking

After searching equivalent groups of labels by using the previous rules, the decision making procedure checks the shifts of their beginning and end boundaries to determine if both groups are definitively equivalent, consequently deemed as correctly labelled.

The user can define a maximum allowed shift according to the context of the group, that is the left context of the first phoneme of the group and the right context of the last phoneme of the group.

For instance, the maximum allowed shift of the end boundary of a vowel will be shorter if it is followed by a nasal consonant than by a liquid consonant (/bar/ vs. /man/)

3.3.6. Conclusion

We have designed a labelling assessment tool which compares simultaneously the transcription and the alignment provided by two labellers. This tool is customisable because the user can define phonological and phonetic rules, specifying the allowed differences between the two labelling results.

Figure 4 shows the results of this labelling assessment method for the sentence *"il y a beaucoup de bouddhistes"*. The labelling results (labels and boundaries) provided by both labellers are displayed under the spectrogram. The units surrounded by a solid line are the groups of labels which are found as equivalent by the whole comparison process. Thus, the units surrounded by dotted line are regarded as mislabelled.

It can be noticed that the following comparison rules have been applied:

[* ə => *]
[j a => i a]

3.4. Labelling assessment with one labeller

3.4.1. Principle When the user has only one available labeller, he can try to shorten the step of manual verification by using the alignment scores given by the automatic labeller.

The main idea is to compute the histogram of the alignment scores for all the sentences of the corpus; then to approximate it by a normal distribution; and finally to consider the sentences corresponding to the ends of the histogram as mislabelled, that is the sentences which score verifies:

$$(\mu - score)^2/\sigma^2 \quad > \quad k$$

3.4.2. Application

We have applied this methodology to label the Swiss French POLYPHONE database. This database contains more than 45000 sentences uttered by 4500 speakers, recorded over the telephone by the SWISS TELECOM PTT and the IDIAP laboratory. The speech files are coded in A-law (8 bits, 8 kHz).

According to the generic methodology explained in section 3, our semi-automatic labelling tool respect the following iterative process:

1. Training of our labelling tool (see section 2) on an already labelled corpus but recorded with very different conditions (16 bits, 16 kHz, high quality microphone, quiet environment). In order to minimise the mismatch between the acquisition conditions, the data have been down-sampled and filtered.

2. Generation of the phonetic graph for every sentence from the orthographic transcription.

3. Alignment of these phonetic graphs.

4. Computation of the histograms of the alignment scores.

5. Selection of the sentences which alignment scores are close to the center of the histogram.

6. Re-training our labelling tool on these selected labelled sentences.

7. Iterating the process from the step 3 while the labelling notably improve.

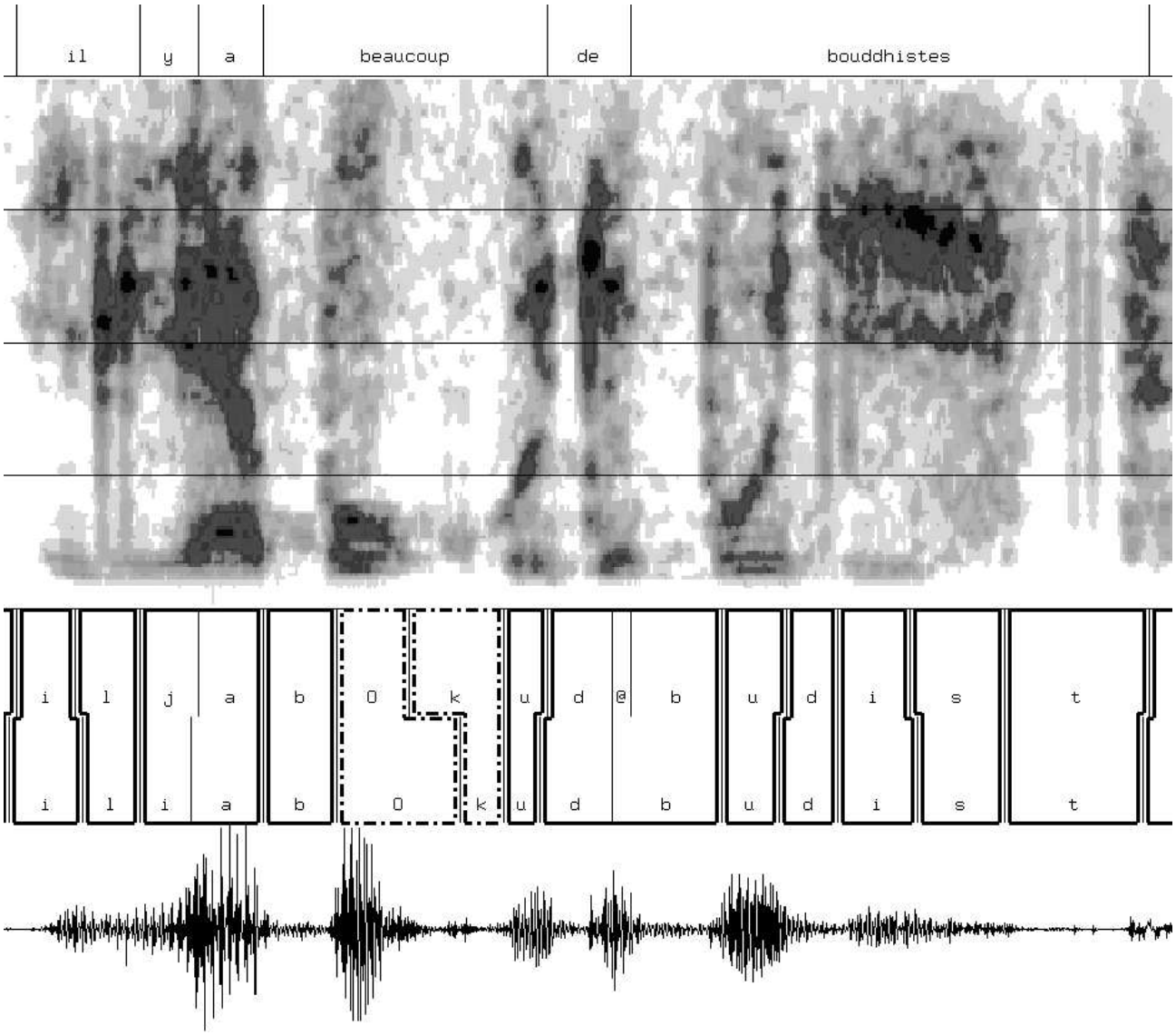

Figure 4: Results of the comparison of two labelling results

8. Finally, the sentences corresponding to the ends of the histogram will have to be manually verified.

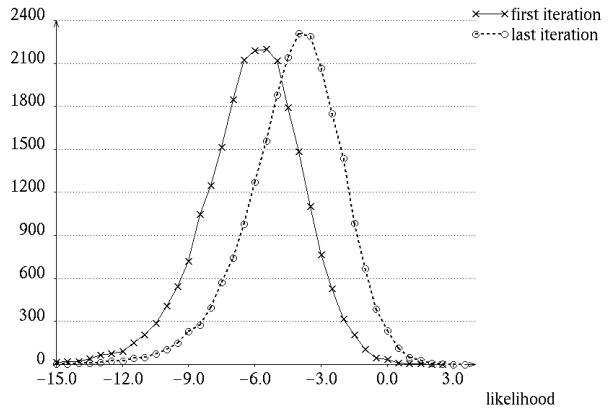

Figure 5: Histogramms

Figure 5 shows two histograms of alignment scores of the first and the fourth iteration passes, for the female speakers of POLYPHONE. It can be noticed that the mean of the scores improves.

4. Conclusion

In this paper, we have first proposed a generic methodology to automatically label large corpora with statistical labelling tools. We have detailed two implementations of this methodology according to the user has one or two automatic labellers.

The implementation with only one labeller allows the user to automatically discard the sentences with singular alignment scores which often correspond to mislabelled sentences. Nevertheless, this method has two main drawbacks: it is blind and not versatile.

By contrast, the second implementation which implies the availability of two automatic labellers use phonological and phonetic rules specified by the user. Thus, this method is customisable and can be easily adapted to other languages. Moreover, for every labelled sentence, this method does not reject the whole sentence but marks any sequence of labels of the sentence as well-labelled or mislabelled. Thus the well-labelled parts can be used to retrain the automatic labellers and the analysis of the mislabelled speech segments could be used to improve the labellers.

5. References

Depambour, P. & Andre-Obrecht, R. & Delyon, B. (1997). On The Use Of Phone Duration And Segmental Processing To Label Speech Signal.In Proceedings of EUROSPEECH'97, Volume 3 pp. 1627-1630.

Eskenazi, M. & Hogan, C. & Allen, J. & Frederking, R. (1997). Issues In Database Creation: Recording New Populations, Faster And Better Labelling. In Proceedings of EUROSPEECH'97, Volume 4 pp. 1699-1702.

Mari, J.-F. & Fohr, D. &Junqua, J.-C. (1996). 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 .

Vorstermans, A. & Martens, J.P. & Van Coile B. (1996). Automatic segmentation and labelling of multi-lingual speech data. in Speech Communication, Vol. 19, No. 4, October 1996, pp. 271-293.