AUTOMATIC ALIGNMENT OF MAP TASK DIALOGS USING WFSTS

D. Caseiro, F.M. Silva, I.Trancoso

INESC-ID/IST
Rua Alves Redol 9, 1000-029 Lisbon, Portugal

C. Viana[*]

CLUL
Av. Prof. Gama Pinto 2, Lisbon - Portugal

ABSTRACT

The goal of this work is the automatic alignment of a map task dialog corpus collected for European Portuguese. The Coral corpus has been orthographically labeled, however off-the-shelf alignment techniques do not work because of the large amount of cross-talk and pronunciation variation. This paper addresses these two issues. The cross-talk problem is dealt with by using a pre-processing stage of channel separation, which proved specially advantageous in the alignment of overlapping speaker turns. The pronunciation variation problem was addressed by including alternative pronunciation rules in the alignment procedure. The alignment is based on *WFSTs* in the sense that its search space is defined by a distribution-to-word (or distribution-to-phone) transducer. Despite many limitations, such as the inadequacy of our current acoustic phone models in terms of voice quality changes (such as laughing), the aligner proved sufficiently robust and demonstrated the feasibility of our alternative pronunciation rules implementation.

1. INTRODUCTION

This work is motivated by the pressing need to have automatic phonetic alignment of spontaneous speech corpora in European Portuguese. Our first experience with this type of material occurred in the framework of the Coral [1] project, in which we collected a corpus of spontaneous dialogs. The corpus was orthographically transcribed, annotating several phenomena of particular importance to modeling the pronunciation variance. However, due to limited funding, it was never aligned at a phonetic level, and even word-level alignment was only done for a very small subset. Hence, automatic alignment becomes crucial. There are, however, two reasons why off-the-shelf alignment techniques do not work: the large amount of cross-talk and pronunciation variation. This paper addresses these two issues.

The cross-talk problem is specially important in overlapping turns. We tried to decrease the observed cross-talk between the two channels, by using source separation techniques, as a pre-processing stage, before doing the alignment.

The alignment procedure uses a *WFST* (Weighted Finite State Transducer) framework, following the promising results we have recently obtained in the context of digital spoken books [2] - in fact, the alignment of a 2h15m audio file ran at 0.03 xRT (excluding acoustic modeling). Our goal here was to test the same type of method with spontaneous speech, in a dialog context, rather than with read speech. In particular, we want to test the feasibility of an extensive set of alternative pronunciation rules to cope with the large pronunciation variability we may observe in this corpus. The issue of language modeling in spontaneous speech will not be addressed here.

This paper thus has 4 main parts, described in the following sections: section 2 describes the Coral corpus, its annotation and some relevant statistics; section 3 is devoted to the channel separation procedure; section 4 describes our aligner and the implementation of alternative pronunciation rules; experimental results are shown in section 5; finally, section 6 summarizes the main conclusions of this work.

2. THE CORAL CORPUS

Coral is a map task dialog corpus, involving spontaneous conversations between pairs of speakers about map directions. It was collected in the framework of a national project sponsored by the PRAXIS XXI program, by a consortium formed by INESC, CLUL, FLUL and FCSH-UNL. In the 16 different pairs of maps, the names of the landmarks were chosen to allow the study of some connected speech phenomena: sequences with /l/ favoring or not its velarization (e.g. *sala malva, sal amargo*); sequences with /s/ in word final position followed by another coronal fricative (e.g. *poços secos*); sequences of plosives formed across word boundaries (e.g. *clube de tiro*); and sequences of obstruents formed within and across word boundaries (e.g. *bairros degradados*).

The recordings involved 32 speakers (students from the Lisbon area), and took place in a small sound proof room at INESC. The two speakers were separated by a distance of about one meter with a small screen wall in between

[*] Names in alphabetical order.

them, whose goal was to avoid direct visual contact between the participants, but did not provide acoustic isolation. The speakers wore close-talking microphones and the recordings were made in stereo directly to DAT and later down-sampled to 16 kHz per channel.

All dialogs were orthographically transcribed following the same transliteration conventions using SGML format of other map task corpora[1]. Audio samples and corresponding transcription of some turns of the pilot dialog that was first recorded can be found in the group's website. [2]

2.1. Corpus analysis

On average in the 64 dialogs, the number of turns was close to 150, although the variation was quite large: from 43 (between two twins) to 305. The lexicon included 2775 different forms (253 of which are abandoned forms which have not been completely pronounced). Altogether, the corpus includes 61181 words. The percentage of turns with overlapping marks is quite high (69%). The number of macro-annotations, denoting some type of noise or voice-quality change was 5607. The majority corresponds to vocal (64%) and intermitent noises (20%), but there is also a significant number of tags annotating laughs or laughing quality (10%). The number of micro-annotations is much higher: 21415, but half correspond to pauses. Table 1 shows the relative percentage of the most significant micro-annotations.

Tag	Description	%
ab	abandoned	2.7
br	broken	1.8
ci	cited items	1.6
ct	contraction	10.2
fp	filled pause	4.3
gg	grunt	5.8
ip	initial partial	5.0
ph	phonetic	7.6
pi	pause with inhalation	6.0
pp	pause	51.1
rp	repeated	3.6

Table 1. Micro-annotations in the Coral corpus.

It is interesting to notice that 56% of the repeated items involve function words and auxiliary verb forms (mostly monosyllabic). As expected, all cited items involve place names with multiple word selected for the elements of the map. Concerning abandoned forms, in 26% of the annotated cases, only a single sound is produced (e.g [s]), and in 36% of the cases, only the first syllable. Broken forms are much less common.

Of particular importance to this work is the analysis of the other micro-annotations which explicitly indicated the adopted pronunciation (phonetic, contraction and initial partial). Altogether, there were 4076 such tags, although we observed some inconsistency among the annotators, who frequently confused them. 31% mark monophthongization with or without vowel coalescence (e.g. [b"aSu] instead of [b"ajSu], for the word *baixo*, meaning down). Another significant percentage (30%) occurs with the prepositions *para* and *por* and their contractions with other words (e.g. [pO] instead of [p6r6 u]); an interesting case is the word *por*, which was only pronounced in its canonical form [pur] 61% of the times, an alternative form [pru] being also very frequent (32%). Truncation of the initial syllable is also fairly frequent (21% in forms of the verb to be *estar*, and 5% in other words). 5% simply mark lengthening vowels or consonants in monosyllabic function words.

3. CHANNEL SEPARATION

Given the recording conditions, a reasonable amount of cross-talk from the other channel is clearly audible. The recordings took place in a controlled environment, but equivalent setups may be found in real applications, where each speaker will likely have a microphone providing a main reference signal, but including a reasonable amount of cross-talk from other audio sources in the same acoustic environment. Several different sources of cross-talk can be identified in the designated setup. In first place, a small direct cross-talk path may exist in the mixing console. Secondly, the direct physical path from the head-mounted phones (where both the primary and secondary signals are reproduced) and the associated microphone. Finally, the normal acoustic path through the studio room from each speaker to the microphone of the other speaker. While the studio walls are made of acoustic absorbing material, a small amount of reverberation may be expected. It is difficult to define a single figure for SNR, which is highly variable during different dialog phases and speaker fluctuations. A rough estimate obtained from the average power measured during speech and silence in each channel yields an SNR during overlapping turns of about 12dB (for the weakest signal) and 25dB (for the strongest one).

Techniques based on independent component analysis (ICA) [3] seemed good candidates for source separation in co-channel speech. While ICA was first formulated for instantaneous mixtures, several extensions were proposed for the separation of convolutive mixtures [4, 5]. However, all these methods assume a constant source signal flow in each mixture channel. This is hardly the case in real dialog situations, where most of the time only one source is present. The overlapping periods are usually short, difficult to detect, and the underlying statistics are not enough to provide

[1] http://www.hcrc.ed.ac.uk/dialogue/maptask.html

[2] http://www.l2f.inesc-id.pt/projects/coral/ortograf.html

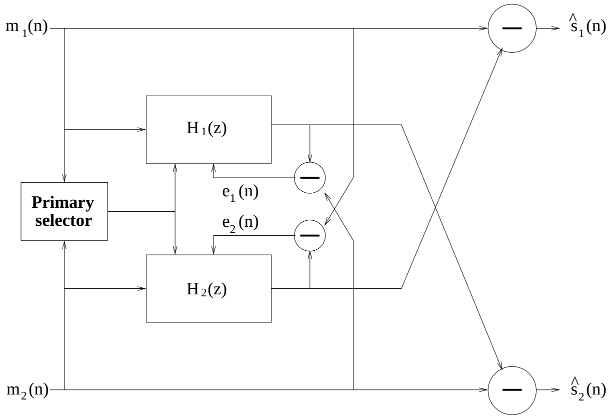

Fig. 1. Symmetric adaptive canceling architecture. $m_1(n)$ and $m_2(n)$ denote the (noisy) mixtures, $\hat{s}_1(n)$ and $\hat{s}_2(n)$ the estimated sources.

reliable independent component analysis.

Given these limitations, a simpler approach was adopted, using an adaptive noise canceling scheme [6], in a symmetric architecture, in order to estimate both source signals simultaneously (Fig. 1). The goal of each adaptive filter is to estimate the cross-talk component in each mixture signal, given the main interfering signal. Each filter had 256 taps (16ms) and was adapted using the standard LMS algorithm.

When both filters are adapted simultaneously, this architecture provides decorrelated but not necessarily separated signals. Moreover, it is prune to stability problems [7]. In order to avoid this limitation, only one of the filters was adapted at each time step. The adaptation decision was made by comparing the short-time energy (1ms) of the two mixture signals, and the signal with larger energy was selected as the primary signal. No attempt was made to avoid adaptation during overlap periods, since these are usually short and not enough to jeopardize the filter estimates. Using this scheme, an average cross-talk reduction of 10dB was achieved for the weaker interference signal, and of 18dB for the stronger interference. The resulting SNR increased to about 30dB and 35db, respectively. At the perceptual level, the interfering signal becomes almost inaudible. No loss of quality was observed in the reference primary signal.

4. ALIGNMENT

Our aligner is based on *WFSTs* in the sense that its search space is defined by a distribution-to-word (or distribution-to-phone) transducer that is built outside the decoder. For the alignment task, that search space is usually build as $H \circ L \circ W$, where H is the phone topology, L is the lexicon and W is the sequence of words that constitutes the orthographic transcription of the utterance. As no restrictions are placed on the construction of the search space, it can easily integrate other sources of knowledge, and can be optimized and replaced by an optimal equivalent one.

In order to cope with possible de-synchronizations between the input and output labels of the *WFST*, the decoder was extended to deal with special input labels that are internally treated as epsilon labels (similar to skip arcs in Hidden Markov Models), but are used to mark time transitions or boundaries. Whenever such end-of-segment labels are crossed, the time is recorded in the current hypothesis. The user may choose to place those labels at the end of each phone *WFST* or at the end of each word *WFST*, depending on choosing either phone-level alignment or word-level alignment, respectively.

4.1. Using phonological rules for phone-level alignment

In doing the alignment, instead of building a lexicon with multiple pronunciations per word, we opted for using phonological rules together with a lexicon of canonical forms, in order to account for alternative pronunciations.

These rules are specified using a finite-state grammar whose syntax is similar to the Backus-Naur-form augmented with regular expressions. We added the operator $\rightarrow$, simple transduction, to the usual set of operators, such that ($a \rightarrow b$) means that the terminal symbol a is transformed into the terminal symbol b. The language allows the definition of non-terminal symbols (e.g. *$vowel*). All rules are optional, and are compiled into *WFSTs*. We do not apply the rules one by one on a cascade of compositions, but, because they are optional rules, we rather build their union in order to avoid the exaggerated growth of the resulting transducer, which can be exponential with the length of the composition cascade. The rule transducer R is thus build as $R = \bigcup_i \Sigma^*(R_i\Sigma^*)^*$ where R_i is the tranducer corresponding to a particular rule expecification expression. The rules are applied as $H \circ R^{-1} \circ R^{-1} \circ R^{-1} \circ L \circ W$, where R^{-1} is the inverse of the rule transducer. The rule transducer R is used three times in order to reduce the dependency on the order of the rules. An example of a sandhi rule specification is:

```
$V = $Vowel | $NasalVow | $Glide | $NasalGli;
DEF_RULE S_z, ($V (S -> z) WORD_BREAK $V)
```

5. EXPERIMENTAL RESULTS

The experimental results described in this section were obtained with acoustic models trained for a broadcast news recognition task [8]. The models use a topology where context-independent phone posterior probabilities are estimated by three MLPs (Multi-Layer Perceptrons) given the acoustic data at each frame, and later combined. The resulting network has 39 output units corresponding to the 38 phones for European Portuguese plus silence. These models cannot yet adequately model for instance laughs and certain filled pauses which are so frequent in the Coral dialog corpus.

The performance of our *WFST*-based aligner has been previously tested at a phone level in the context of a small manually labeled read speech corpus [2], using a much more limited set of rules, but not for spontaneous speech. In the present context, only word level tests can be done, since we only have one pilot dialog manually annotated with time stamps for word boundaries, and not for phone boundaries. We started by measuring the average absolute error between the reference time stamps and the automatic ones for each word start, without using either channel separation or alternative pronunciation rules. The lexicon, which we shall denote by Lex0, includes only canonical forms. Multiple pronunciations are exclusively used for heterophonic homographs (amounting to 21).

For the left channel, corresponding to the speaker designated as *Giver*, the average error was 0.380s. For the right channel, corresponding to the speaker designated as *Follower*, the average error was 2.346s (first line of table 2). The larger errors obtained with the *Follower* can perhaps be due to much smaller turns, many of them grunts largely overlapping with the *Giver*'s turns. In fact, whereas the *Giver* spoke 674 words during these approximately 5 minutes, the *Follower* spoke only 409 words. Without using channel separation, we observe that the end of the turn is not properly detected, which causes words from one of the speakers to be frequently aligned during the other speaker's turn. The problem is aggravated when overlap occurs.

When channel separation is used, the average error decreases as shown in the second line of the same table. The alignment obtained with the separated signals is fairly good. An analysis of the largest errors shows they may be due to the fact that we did not try to align laughs, which causes severe misalignments in the neighboring words. The performance in overlapping turns is on the same level as the one in non overlapping turns.

Next we investigated the relevance of providing alternative pronunciations for function words and forms of the verb *estar*, which were so frequently marked with micro-annotations in our corpus. The values obtained with this new lexicon (Lex1, including multiple pronunciations for 40 forms) are shown in the third and fourth lines of table 2, without and with channel separation respectively. Given these results, further tests with alternative pronunciation rules were done only with Lex1 and channel separation.

The main phonological aspects that alternative pronunciation rules are intended to cover are: (1) intra-word vowel devoicing; (2) voicing assimilation; and (3) vowel and consonant deletion and coalescence. Both (2) and (3) may occur within and across word boundaries. Some common contractions are also accounted for, with both partial or full syllable truncation and vowel coalescence. Vowel reduction, including quality change, devoicing and deletion, is specially important for European Portuguese, being one of the features that distinguishes it from Brazilian Portuguese and that makes it more difficult to learn for a foreign speaker. As a result of vowel deletion, rather complex consonant clusters can be formed across word boundaries, as mentioned in section 2. Even simple cases, such as the coalescence of the two plosives in *grade de ferro*, raise interesting problems of whether they may be adequately modeled by a single acoustic model for /d/.

The results obtained with 56 rules are shown in the last line of the table. Phone alignment error would be a much more adequate measure, but unfortunately, we do not yet have reference labels. We observed that the misalignments due to the absence of models of laughs, although affecting only the neighboring words, can be as large as 5s, and almost destroy any potential improvements brought by the use of the rules. For the left channel, for instance, only 1.7% of the word boundary errors are above 1s and most of these errors are due to such segments. The next step will clearly be marking them and creating adequate acoustic models. In order to do this for the whole corpus, automatic alignment followed by posterior manual correction is crucial. Whereas the alignment obtained with the original signals without channel separation is too bad to serve as a starting point, the one obtained with channel separation seems good enough.

Av. error [s]	Ch. sep.	Left-ch.	Right-ch.
Lex0 / no rules		0.380	2.346
Lex0 / no rules	√	0.097	0.151
Lex1 / no rules		0.343	2.334
Lex1 / no rules	√	0.086	0.146
Lex1 / rules	√	0.077	0.143

Table 2. Average word alignment error.

Figure 2 illustrates the performance of the aligner in the presence of an overlapping turn. The top section shows the original signals without separation (right channel above left channel). The middle section shows the corresponding signals after channel separation, where the reduced cross-talk can be observed. The bottom section shows three sets of labels corresponding to the left channel only: the top one is the manual reference labeling; the middle one is the automatic labeling, obtained without channel separation (Lex0 / no rules); the bottom one is the automatic labeling obtained with channel separation (Lex1 / rules). The better match with the manual labeling may be observed, as the aligner does not try to match the word *ferro* with the word *sim* (yes) spoken by the other speaker, as it did without channel separation. Notice, however, that the coalescence of the two plosives that occurred in *grade de ferro* was not marked by the manual annotator either. The figure also illustrates the difficulties of assessing the quality of automatic labelling without having reference phone-level labels.

6. CONCLUSIONS AND FUTURE WORK

This paper described our first steps toward the study of spontaneous speech in dialogs. We started by characterizing our corpus and explaining how we used channel separation for dealing with the stereo recordings. Our *WFST*-based aligner proved sufficiently robust to be able to process fairly long dialogs with overlapping turns, despite many limitations, namely in terms of the absence of models for voice quality changes that are so frequent in this corpus.

The automatic alignment of this corpus is really a crucial step for the recognition of spontaneous speech in the context not only of Coral but also of another project dealing with Broadcast News. Our current recognition error rates (18.8% for read speech vs. 40.6% for spontaneous speech) further increase our motivation to have an annotated corpus that would allow us to properly address pronunciation modeling problems in spontaneous speech.

It is also worth noting that our initial acoustic models were created from a very small amount of manually anotated read speech. Retraining was done in a bootstrapped process using large amounts of automatically annotated read speech. We believe that some of the pronunciation variation was included in these models, as the posterior probabilities were estimated using a context of 7 frames (3 to the left and 3 to the right of the center frame). Retraining with speech material that is automatically aligned using alternative pronunciation rules (and manually verified) is thus one of the tasks that we are planning for the near future.

7. ACKNOWLEDGEMENTS

The authors would like to thank the two anonymous reviewers for many helpful suggestions and also our colleagues Hugo Meinedo, João Paulo Neto and Isabel Mascarenhas for their prompt cooperation. This work was partly done in the context of European Project BLISS (Blind Source Separation and Applications - IST-1999-14190). The present work is also part of Diamantino Caseiro's PhD thesis, initially sponsored by a FCT scholarship. INESC-ID Lisboa had support from the POSI program of the "Quadro Comunitario de Apoio III".

8. REFERENCES

[1] I. Trancoso, C. Viana, I. Duarte, and G. Matos, "Corpus de diálogo coral," in *Proc. PROPOR'98 - III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, Porto Alegre, Brazil, November 1998.

[2] D. Caseiro, H. Meinedo, A. Serralheiro, I. Trancoso, and J. Neto, "Using wfsts for aligning spoken books," in *Proc. HLT 2002 - Human Language Technology Conference*, San Diego, California, March 2002.

[3] J. Herault and C. Jutten, "Blind separation of sources, part I, an adaptive algorithm based on neuromimetic architecture.," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[4] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[5] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, March 2000.

[6] B. Widrow, J. Glover, J. McCool, C. Williams, R. Hearn, J. Zeidler, Dong E., and R. Goodlin, "Adaptive noice canceling: Principles and applications," *Proceedings of the IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.

[7] S. Gerven and D. Compernolle., "Signal separation by symmetric adaptive decorrelation: stability, convergence and uniqueness," *IEEE Transactions on Signal Processing*, vol. 43, no. 7, pp. 1602–1612, July 1995.

[8] H. Meinedo and J. Neto, "Combination of acoustic models in continuous speech recognition hybrid systems," in *Proc. ICSLP'2000 - Internation Conference on Speech and Language Processing*, Beijing, China, October 2000.

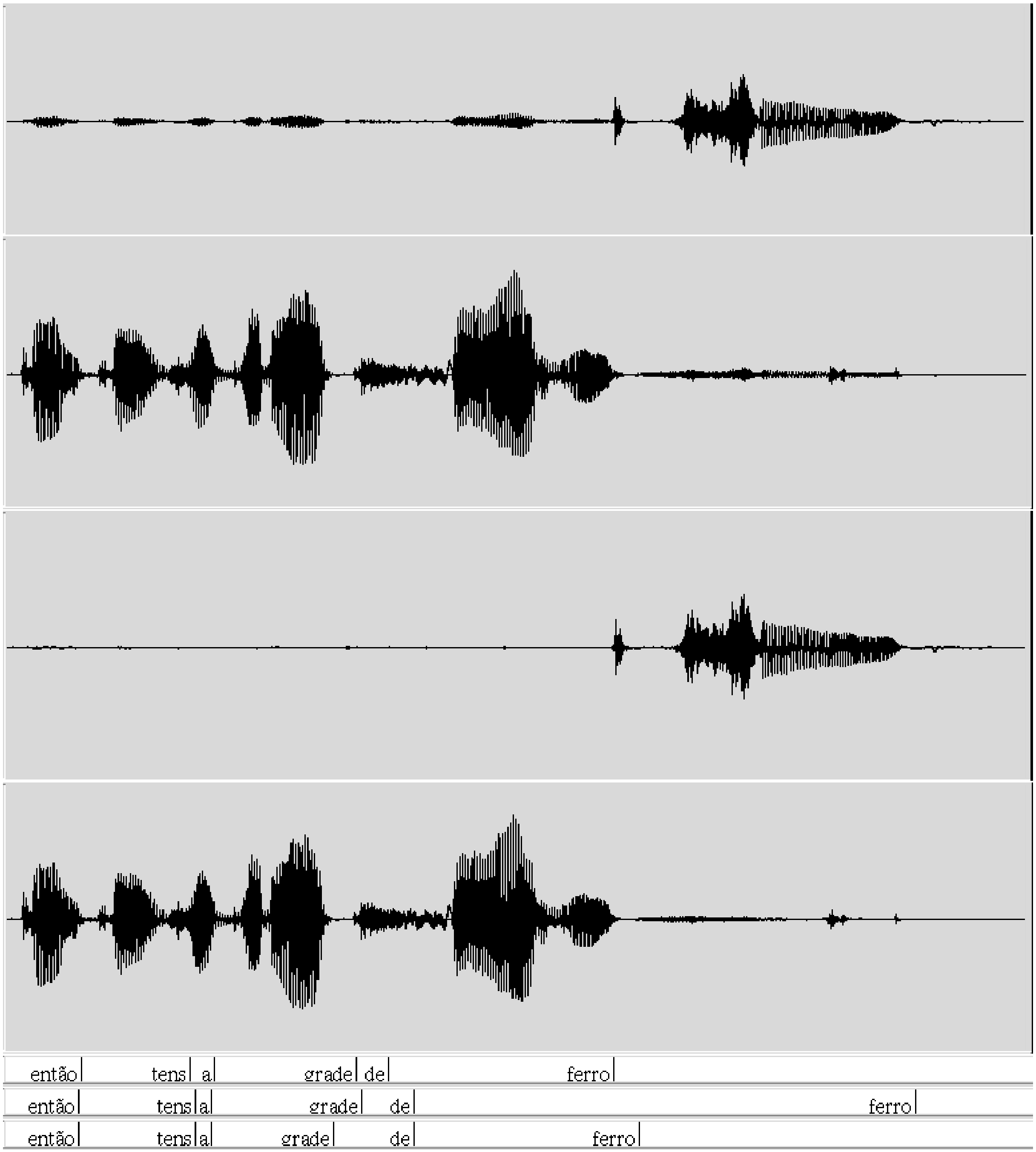

Fig. 2. Aligner performance without and with channel separation. Waveforms (from top to bottom): right channel (without source separation); left channel (without source separation); right channel (with source separation); left channel (with source separation). Orthographic labels (from top to bottom - left channel): manual reference labels; automatic labels (without source separation); automatic labels (with source separation).