CoALT: A Software for Comparing Automatic Labelling Tools

Dominique Fohr, Odile Mella
LORIA-INRIA Nancy Grand Est
Bâtiment C LORIA, B.P. 239
F54506 Vandoeuvre-lès-Nancy, France
E-mail: {fohr,mella}@loria.fr

Abstract

Speech-text alignment tools are frequently used in speech technology and research. In this paper, we propose a GPL software CoALT (Comparing Automatic Labelling Tools) for comparing two automatic labellers or two speech-text alignment tools, ranking them and displaying statistics about their differences. The main feature of CoALT is that a user can define its own criteria for evaluating and comparing the speech-text alignment tools since the required quality for labelling depends on the targeted application. Beyond ranking, our tool provides useful statistics for each labeller and above all about their differences and can emphasize the drawbacks and advantages of each labeller. We have applied our software for the French and English languages but it can be used for another language by simply defining the list of the phonetic symbols and optionally a set of phonetic rules. In this paper we present the usage of the software for comparing two automatic labellers on the corpus TIMIT. Moreover, as automatic labelling tools are configurable (number of GMMs, phonetic lexicon, acoustic parameterisation), we then present how CoALT allows to determine the best parameters for our automatic labelling tool.

Keywords: speech-text alignment**,** automatic labelling, speech processing

1. Introduction

Speech-text alignment tools are frequently used in speech technology and research: for instance, for training or assessing of speech recognition systems, the extraction of speech units in speech synthesis or in foreign language learning. We propose the software CoALT (Comparing Automatic Labelling Tools) for comparing two automatic labellers or two speech-text alignment tools, ranking them, and displaying statistics about their differences. Our software will be available under GPL.

The main feature of our software is that a user can define its own criteria for evaluating and comparing two speech-text alignment tools. These criteria may vary depending on the application task. For training speech recognition systems, the most important is to find the exact sequence of phonemes, but phoneme boundaries are of little significance due to embedded training. By contrast, in foreign language learning, speech-text alignment must find accurate boundaries in order to compute acoustic or prosodic features for guiding the language learner. In post-synchronisation applications, the critical aim is to find very precise boundaries.
With CoALT, a user can give more importance to either phoneme labels or phoneme boundaries. Indeed, the CoALT elastic comparison algorithm takes into account time boundaries.
Moreover, by providing a set of phonetic rules, a user can define the allowed discrepancies between the automatic labelling result and the hand-labelling one.

Another important feature of CoALT is that it accepts that some hand-labelled boundaries are fuzzy, that is, the human labeller does not have criteria to place the boundary, for instance when /a/ is followed by /R/ in the same French syllable. CoALT doesn't take into account shifts on fuzzy boundaries.

Beyond ranking, our tool provides useful statistics about each labeller and above all about their differences and can highlight the drawbacks and advantages of each labeller.
Of course, CoALT can also be used to compute statistics on only one automatic labeller but it was not designed for that since its aims are comparing and ranking two automatic labellers. The next version will handle any number of automatic labellers.

Another use of CoALT is the tuning of an automatic labeller. Often automatic labelling tools are configurable (number of GMMs, phonetic lexicon, acoustic parameterisation), by comparing them, our software allows to determine the best parameters according to a given task.

We have evaluated our software for the French and English languages but it can be used for another language by simply defining the list of the phonetic symbols and optionally a set of phonetic rules. In this paper we present results for the English language.

The next section describes the CoALT software. We then present an example of its usage for comparing two automatic labellings of the TIMIT test corpus. In section 4, we use CoALT to determine the best parameters of our automatic labelling tool.

2. CoALT Description

This tool compares the results of two automatic labelling tools to a reference manual labelling in order to rank them. The ranking is based on the computation of insertions, deletions, substitutions, and shift between boundaries. Besides ranking, the tool provides information about errors made by each automatic labelling tool and emphasizes their differences (for instance insertion, deletion).

2.1 Architecture

CoALT consists of seven steps:

- the conversion of the input labels,
- the application of equivalence rules,
- the alignment of the results of each automatic labeller with the manual labelling,
- the application of phonological and acoustic-phonetic rules,
- the merging of the results of the two previous alignments,
- the scoring and ranking of the labellers,
- the extraction of statistics about the two automatic labellers.

2.1.1 Conversion of the input labels

Given A1 (resp. A2) the result of the alignment of a sentence by the first (resp. second) automatic labelling tool, and Ref the hand-labelling of the same sentence.
As the two labelling tools and the human labeller may use different sets of phonetic symbols, the user must define a common phonetic alphabet. He must also define three sets of rewriting rules used by CoALT to convert the labels of A1, A2 and Ref into the labels of the common alphabet.
Examples:

- the manual labelling of TIMIT makes a distinction between the closure and the burst of a plosive but usually automatic labellers don't do it : [tcl t ⇒t],
- automatic labellers don't manage the IPA : [ɛ⇒E] for French language.

2.1.2 Application of equivalence rules

If the user does not want to make a distinction between two phonemes or allophones, he can define the corresponding equivalence rules. Before alignment, A1, A2 and Ref are rewritten using these rules.
For instance, if the user is not interested in the difference between the two phonemes /e/ and /E/ in French or /ax/ and /ah/ in English, the rules will be:

- for French: [e ⇒ E]
- for English: [ax ⇒ ah]

2.1.3 Alignment of the results of each automatic labeller

In this step, we first perform two alignments: between A1 and Ref and between A2 and Ref. The alignment algorithm is based on an elastic comparison algorithm (DTW) and takes into account the labels and their time boundaries. The user can configure the algorithm by setting:

- *w* and w_t, weighting coefficients so that the algorithm favours either the matching of the labels or the closeness of the boundaries. This method is different from that implemented by Dobrisek & Mihelic (2011);
- *ins[p]*, an insertion penalty for every phoneme; phonemes with a weak penalty can be inserted more easily. For instance, the user can assign a weaker insertion penalty for extra speech segments to help the algorithm to consider them as inserted labels;
- *del[p]*, a deletion penalty for every phoneme; phonemes with a weak penalty can be deleted more easily. For instance, in French, the mid-central vowel schwa can be omitted, the user can assign to it a weaker deletion penalty;
- *sub[p,q]*, a substitution matrix including the penalties for making a substitution between two phonemes. For instance, to avoid that the algorithm matches a vowel with a stop consonant, the user can define a larger penalty for the substitution of a vowel by a stop consonant than for the substitution of two vowels.

The following equations define the distance *d* between the reference labelling (r_i, $i=1,...,n$) and the automatic labelling (a_j, $j=1,...,m$).

$$d_{i,j} = \min \begin{cases} d_{i-1,j-1} + w.sub[r_i, a_j] + w_t.sub_t(r_i, a_j) \\ d_{i-1,j} + w.del[r_i] + w_t.del_t(r_i) \\ d_{i,j-1} + w.ins[a_j] + w_t.ins_t(a_j) \end{cases}$$

Where $subt(r_i,a_j)$ is the Manhattan distance between beginnings and ends of the two phonemes ; del_t and ins_t are monotonically increasing functions of the phoneme duration.

2.1.4 Application of phonological and acoustic-phonetic rules

To be general and to fit the requirements of the user, CoALT needs to know the degree of similarity desired by the user between the automatic and manual labelling. More precisely, when the different labelling tools do not have the same level of acoustic-phonetic accuracy or when the user wants some differences to be ignored, the user can define rules to allow some substitutions (described as *allowed* further), insertions, and deletions. The rules are applied on the results of the two alignments computed at the previous step. Whenever CoALT applies a rule, it stores this information.
Some examples of rules are presented below.

Examples of allowed substitutions

If the user does not want to rank the two automatic labellers according to the difference between the two phonemes /e/ and /E/ in French or /ax/ and /ah/ in English, he can set:

- for French: [e ⇒ E] [E ⇒ e]
- for English: [ax ⇒ ah] [ah ⇒ ax]

We can notice that using equivalence rules is different from using allowed substitution rules. As equivalence rules are applied before the alignment algorithm and the phonological rules after, the final alignment results can differ. Moreover, the advantage of using phonological rules is that the user can know when and how often they have been applied. CoALT displays the allowed substitutions applied and their number (cf. Table 7 in section 3.4.4).

Examples of allowed deletions

- Automatic labelling tools seldom model the glottal stop:
  [q $\Rightarrow$ ∅], ∅ denotes the null symbol.
- Some words have different pronunciations according to the phonetic dictionary:
  [i j a $\Rightarrow$ i ∅ a].

Examples of allowed insertions

- When two successive words start and end with the same consonant, some speakers utter only one phoneme.
  For instance: "porte tournante": [t t $\Rightarrow$ t]
- Non-speech events (pause, noise, breathing, cough...) are not modelled with the same accuracy between labellers. Some hand-labelling don't discriminate non-speech parts of the audio signal:
  [sil xx $\Rightarrow$ sil ∅], xx denotes a noise.

When a rule is applied on a sequence of at least two phonemes, the internal boundaries of the sequence are tagged as fuzzy in the automatic (A1 or A2) and reference (Ref) labellings. For instance, applying the rule [i j a $\Rightarrow$ i ∅ a], results in three fuzzy boundaries: between /i/and /j/; /j/ and /a/; /i/ and /a/. The shifts are not computed on fuzzy boundaries.

2.1.5 Merging of the two previous alignments

As shown in Figure 1, we then merge both previous alignment results (each automatic labelling with the reference) into a single alignment.
The symbol "*" in the reference means an insertion by at least one automatic labeller.
The symbol "*" in an automatic labelling result means a disallowed deletion and the symbol "+", an allowed deletion.

Alignment between A1 and Ref

Ref	sil	xx	l	ae	m	z	*	ay	v
A1	sil	+	l	ae	m	z	hh	ae	v

Alignment between A2 and Ref

Ref	sil	xx	l	ae	m	z	ay	v
A2	sil	+	l	ae	m	*	ae	v

Merged alignments

Ref	sil	xx	l	ae	m	z	*	ay	v
A1	sil	+	l	ae	m	z	hh	ae	v
A2	sil	+	l	ae	m	*		ae	v

Figure 1: Example of alignment merging for the start of the sentence "lambs have..."; xx denotes noise.

2.1.6 Scoring and ranking of the automatic labellers

After all the sentences of the reference corpus have been processed in the previous steps, an alignment score is computed for each automatic labeller. Therefore they can be ranked. The score of each automatic labeller is based on all the deletions, insertions and substitutions that are not allowed and the boundary shifts.
The boundary shifts are computed for the matching phonemes if they are identical or if the substitution is allowed by a rule. Moreover, in the case of a fuzzy boundary, no shift is computed. We compute the total number of boundary shifts (TNBS) that are greater than a threshold defined by the user.
The final score of each automatic labeller is a linear combination of all the deletions, insertions and substitutions that are not allowed and the TNBS. The user chooses the linear weights according to the importance he gives to the type of errors.

Moreover, CoALT takes into account the systematic bias that can occur between the reference boundaries and the boundaries given by an automatic labeller. For instance, a labelling tool based on HTK puts the beginning of a phoneme at the beginning of the analysis window. For that, CoALT performs two passes. In the first one it computes the average shift between the boundaries of all phonemes that have been mapped by the alignment algorithm. This average shift is then subtracted from all the boundaries provided by the automatic labeller. In the second pass, coALT again performs all the steps including the alignment.

We can notice that the computed score of each automatic labeller is not an absolute score. Indeed, the application of a rule between one of the automatic labelling and the reference labelling can generate fuzzy boundaries in the reference labelling. Because the user can choose its own evaluating criteria by defining rules, it would be unfair to compare independently each automatic labeller with the reference labelling. Indeed, if the comparisons are done independently, the fuzzy boundaries generated in the reference labelling could be different. Therefore, the number of boundaries which are taken into account to compute the shifts will be different. This is why CoALT compares the automatic labellers together. When a reference boundary is set as fuzzy for one automatic labeller, it is also considered as fuzzy for the other even if no rule has been applied.

2.1.7 Extraction of statistics

In this last step, CoALT displays for each automatic labeller:

- the *n* phonemes which were most inserted (disallowed insertions) ,
- the *n* phonemes which were most deleted (disallowed deletions),
- the *n* couples of phonemes which were most confused (disallowed),
- the average shift of the beginning boundary per phoneme and per left context,
- the average shift of the end boundary per phoneme and per right context,

- the *n* phonemes which were most inserted (allowed insertions) ,
- the *n* phonemes which were most deleted (allowed deletions),
- the *n* couples of phonemes which were most confused (allowed).

Some statistics are also computed per class of phonemes and per class of contexts defined by the user.

Above all, the tool extracts from the statistics the most relevant differences between the two automatic labellers. This helps to highlight the drawbacks and advantages of each labeller. In the next section, Table 2 shows an example of the statistics displayed by CoALT.

3. An example

We present in this section an example of using of CoALT to compare two automatic labellings of the TIMIT Test data. The reference labelling is the hand-labelling provided with the TIMIT database.
Both automatic labellers L1 and L2 are based on HMM acoustic models and on MFCC (Mel Frequency Cepstral Coefficient) parameterisation with a 10ms frame shift. The acoustic models were trained on the TIMIT Train corpus. For the labeller L1, the training stage uses the sequence of phonemes given by the manual labelling of the sentence. For L2, only the sequence of words was used and the phonetic transcriptions of every word were extracted from the CMU Pronouncing Dictionary v.0.6. The phonetic alphabet is the TIMITbet.
As phonetic lexicon, L2 uses only the CMU Pronouncing Dictionary v.0.6, while L1 uses in addition some pronunciation variants extracted from the TIMIT Train corpus (more precisely, the pronunciation variants giving a coverage rate of 50%, as detailed section 4.3) .

3.1 The equivalence rules

According to the section 2.1.2, we define the following equivalence rules. These rules are similar to those proposed in (Lee & Hon, 1989):
[ux ⇒ uw]
[axr ⇒ er]
[em ⇒ m]
[en ⇒ n]
[el ⇒ l]
[nx ⇒ n]
[ix ⇒ ih]
[eng⇒ ng]
[hv ⇒ hh]
[ax-h ⇒ ax]

3.2 The phonological rules

According to the section 2.1.3, we define the following phonological rules:
[q ⇒ ∅]
[s# ⇒ ∅]
[dx ⇒ t]
[dx ⇒ d]

3.3 Scoring and ranking

CoALT provides an error score for comparing the two labellers. The lower the score, the better the labeller is. We can notice in Table 1, that L1 is better because it makes less disallowed confusion, and disallowed insertions.

Number of phonemes in the reference labelling	57668	
Labeller	**L1**	**L2**
disallowed insertions	1679 (2.9%)	2316 (4.0%)
disallowed deletions	521 (0.9%)	389 (0.7%)
disallowed substitutions	4625 (8.0%)	6247 (10.8%)
beginning shift > 20ms	9615	9608
end shift > 20ms	9299	9424
score about TNBS	16.4%	16.5%
final score	28.2%	32.0%

Table 1: The final scoring for the two automatic labellers

3.4 Examples of statistics

We present here some statistics computed by CoALT.

3.4.1 Disallowed deletions

Table 2 shows the disallowed deletions found by the elastic comparison algorithm when CoALT compares the results of each automatic labeller and the manual labelling of TIMIT. For every phoneme of the manual labelling, our tool counts the number of deletions made by each automatic labeller (Nb1 and Nb2). It then sorts their difference (Nb1-Nb2) and displays the *n* greatest and the *n* smallest values. The first rows of Table 2 show the phonemes for which L2 gives an improvement; on the other hand the last rows show the phonemes for which L2 degrades. This kind of results can be useful, for instance, when we want to test the impact of changing one acoustic model used by the automatic labeller.

Nb1-Nb2	**Nb1**	**Nb2**	**Label in Ref**
36	65	29	ax
33	150	117	r
24	51	27	ih
16	23	7	d
8	39	31	hh
6	7	1	eh
6	101	95	t
6	16	10	y
....	....	....	...
-1	0	1	ch
-1	3	4	dx
-1	2	3	s
-1	1	2	zh
-2	0	2	jh
-7	9	16	iy

Table 2: Disallowed deletions

CoALT also sorts on Nb1 and Nb2 and the *n* first values are displayed as shown in Table 3. This classical kind of results could be used to highlight that a phone is difficult to detect by one of the labellers.

Labeller 1		Labeller 2	
Nb1	**Label**	**Nb2**	**Label**
150	r	117	r
101	t	95	t
65	ax	31	hh
51	ih	29	ax
39	hh	27	ih
23	d	16	iy
16	y	10	y

Table 3: Top seven phonemes producing disallowed deletions

3.4.2 Allowed deletions

CoAlt performs the same calculations for the deletions allowed by the phonological rules defined in 3.2. According to these rules, allowed deletions can occur for only two phonemes. Therefore, in Table 4, the top *n* reduces to 2 and the user can know how many times the rule was applied by the comparison tool.

Nb1	label	Nb2	label
983	q	1010	q
544	s#	599	s#

Table 4: Top *n* phonemes producing allowed deletions

3.4.3 Insertions

CoALT displays similar tables for the disallowed and allowed insertions made by the two automatic labellers.

3.4.4 Substitutions

In the same manner as for deletions, CoALT computes statistics about disallowed and allowed substitutions. Table 5 shows the disallowed substitutions sorted with respect to the difference (Nb1-Nb2) and Table 6 displayed the top-ten disallowed substitutions for each labeller.

Nb1-Nb2	Nb1	Nb2	Label in Ref	Substituted label
201	201	0	ih	ax
177	292	115	ax	ih
55	82	27	eh	ih
48	48	0	ah	ax
38	114	76	aa	ao
34	55	21	q	s#
25	36	11	ae	eh
25	25	0	ey	ih
22	30	8	ah	ih
.....	.....	.....	....	....
-37	125	162	eh	ae
-43	127	170	ao	aa
-54	62	116	er	ah
-65	15	80	jh	y
-86	43	129	ih	ae
-176	74	250	er	r
-184	12	196	er	uh
-732	457	1189	ax	ah
-877	553	1430	ih	ah

Table 5: Disallowed substitutions

Nb1	Label in ref	Subst. label	Nb2	Label in ref	Subst. label
553	ih	ah	1430	ih	ah
457	ax	ah	1189	ax	ah
292	ax	ih	250	er	r
243	ih	iy	246	ih	iy
201	ih	ax	196	er	uh
169	q	t	170	ao	aa
147	iy	ih	167	q	t
127	ao	aa	162	eh	ae
125	eh	ae	142	iy	ih
120	s	z	130	ih	eh

Table 6: Top ten disallowed substitutions for each labeller

The last rows of Table 5 show that adding pronunciation variants reduce the number of substitution errors.

Table 5 also shows a huge difference between L1 and L2 for the substitution between /ax/ and /ah/. The labeller L2 use the CMU dictionary which does not make a distinction between these two phonemes (*but*: /b ah t/ and *about*: /ah b aw t /).

If the user is not interested in discriminating /ax/ and /ah/, he can add the phonological rule : [ax $\Rightarrow$ah]. In this case, the substitutions between /ax/ and /ah/ will be counted as allowed substitutions (cf. Table 7) but not in the final score. L1 final score improves from 28.2% to 27.5% and L2 final score from 32.0% to 30.2%.

Nb1	Label in ref	Substituted label	Nb2	Label in ref	Substituted label
457	ax	ah	1185	ax	ah
294	dx	t	675	dx	t
188	dx	d	228	dx	d
48	ah	ax			
10	d	dx			
9	t	dx			

Table 7: Top ten allowed substitutions

3.4.5 Boundary shifts above a given threshold

Statistics are computed about boundary shifts over a given threshold defined by the user. For both automatic labellers and for every phoneme of the manual labelling, our tool counts (for every left context) the number of beginning boundaries whose shift is greater than a threshold (Nb1 and Nb2). It then sorts their difference (Nb1-Nb2) and displays the *n* greatest and the *n* smallest values (cf. Table 8).

It also sorts on Nb1 and Nb2 and the n first values are displayed as shown in Table 9.

The software computes the same statistics for the end boundaries.

Nb1-Nb2	Nb1	Nb2	left context	label
65	68	3	ih	er
53	76	23	iy	ih
34	39	5	y	er
33	33	0	ae	jh
23	25	2	jh	er
23	226	203	s#	d
21	132	111	r	iy
20	74	54	ih	z
.....	.....	.....	....	....
-19	5	24	k	y
-19	37	56	w	aa
-23	46	69	n	t
-25	75	100	s	s#
-26	9	35	s#	y
-28	67	95	q	ae
-35	16	51	l	ah
-35	62	97	z	s#

Table 8: Shifts of the beginning boundaries for each pair (left context, phoneme) in the reference labelling

Nb1	left context	label	Nb2	left context	label
226	s#	d	203	s#	d
182	ao	l	182	ao	l
175	aa	r	169	aa	r
153	t	s#	153	q	ao
152	y	ih	151	y	ih
143	q	ao	140	t	s#
134	s#	dh	132	q	ih

Table 9: Top 7 shift of beginning boundaries with corresponding pairs (left context, phoneme)

3.4.6 Average shifts of boundaries per phoneme class

The user can define the class of phoneme and the class of context for which he wants to get the average shift of the beginning (resp. end) boundaries. In this example, we define the following classes:

- voiced stops: /b,d,g/
- unvoiced stops: /p,t,k/
- vowels: /ax,aa,ae,ah,ao,aw,ay,eh,er,ey,ih,iy,ow, oy,uh,uw/
- affricates: /ch, jh/
- fricatives: /dh,f,hh,s,sh,th,v,z,zh/
- glides: /l,m,n,ng,r,w,y/
- silence: /s#/

For the classes of context we gather voiced and unvoiced stops, and affricates and fricatives. Table 10 shows an extract of the results for beginning boundaries.

L1		L2			
Average 1	Nb1	Average 2	Nb2	Left context	Phoneme
39.7 ms	244	40.1 ms	244	Stops	Voiced stops
7.8 ms	1839	8.0 ms	1839	vowels	Voiced stops
16.2 ms	296	25.1 ms	297	fricatives	Voiced stops
11.7 ms	725	13.9 ms	739	glides	Voiced stops
39.0 ms	350	42.5 ms	350	silence	Voiced stops
14.8 ms	3454	16.5 ms	3469	all	Voiced stops

Table 10: Average shifts of beginning boundaries per phoneme and left context classes

4. Tuning of parameters for automatic labelling

4.1 Introduction

We used CoALT to choose the best parameters of our automatic labeller. This automatic labeller is based on HMMs and on a MFCC parameterisation with a 10 ms frame shift. It needs a set of phoneme models and a phonetic lexicon. We chose context-independent models because it has been shown they provide better alignment (Toledano & Gomez, 2003).

For each sentence of the Test part of the TIMIT corpus, the labeller provides the sequence of phonemes and their boundaries.

We have tested the number of pdfs and several phonetic lexicons, and different sets of models,

For all the experiments, the HMMs were trained on the Train part of the TIMIT Corpus and the reference labelling was the manual labelling provided with the TIMIT corpus. Unless otherwise stated, the rules are those presented in sections 3.1 and 3.2.

The shift threshold was set to 20 ms as a good compromise. Indeed, Kawai and Toda (2004) have shown that in Japanese phoneme boundaries put by four human labellers can differ on average up 20ms. Wesenick and Kipp (1996) have compared the labelling of 64 German sentences by three human labellers and an automatic labeller: they have shown that as much as 96% of hand-labelled boundaries are within a range of 20 ms.

4.2 Tuning of the number of pdfs

In this first experiment, the acoustic models were trained using only the sequence of phonemes provided by hand-labelling of the TIMIT Train corpus. The phonetic lexicon was the CMU Pronouncing Dictionary v.0.6. We assessed our labeller with respect to several numbers of pdfs: 1, 2, 4 and 8. Increasing the number of pdfs degrades the quality of the alignment. We have got the same result as that obtained by (Toledano & Gomez, 2003) for 20ms shift. The best results were obtained with 1 or 2 pdfs and

the difference between them is not significant. Table 11 shows that increasing the number of pdfs increases the number of boundary shift errors.
For the following experiments we kept the phoneme models with one pdf.

Number of phonemes in the reference labelling	57668	
Labeller	**with 1 pdf**	**with 8 pdfs**
disallowed insertions	2303 (4.0%)	2307 (4.0%)
disallowed deletions	370 (0.6%)	316 (0.5%)
disallowed substitutions	6198 (10.7%)	6421 (11.1%)
beginning shift > 20ms	9573	10259
end shift > 20ms	9298	9932
score about shifts	16.4%	17.5%
final score	31.7%	33.2%

Table 11: Final scoring for 1 and 8 pdfs

4.3 Influence of the phonetic lexicon

For this second experiment, we investigated the influence of the pronunciations available for each word. As 80% of all the occurrences of the TIMIT Test words belong to the Train corpus, we wanted to evaluate how adding pronunciation variants observed in the manual labelling of the Train corpus influences the quality of the automatic labelling.
The baseline lexicon was extracted from CMU Pronouncing Dictionary v.0.6. It contains 2379 words and 2890 pronunciations (i.e. 1.2 variants per word). We built the other lexicons by adding the pronunciation variants observed in the manual labelling of the Train corpus (Kim et al., 2011). For that, for each word of the Train corpus, we sorted the variants by their number of occurrences. Then, for each word, we selected as many variants as necessary to achieve at least a given coverage of the pronunciations. We chose the following coverage rates: 50%, 75%, 90% and 100% (ALL) which correspond respectively to 7967, 9170, 10355 and 11220 pronunciations. As CoALT compares the labelling tools two by two, Table 12 provides the results achieved by CoALT when it performed pairwise comparison of the five automatic labellers according to the different lexicons.
We can notice that adding frequent pronunciation variants improves the quality of the alignment but adding all the phonetic variants degrades the alignment. Indeed, the less common variants often correspond to atypical pronunciation and add noise to the automatic labeller.

L1/L2		**Labeller L2**			
		+ 50%	+75%	+90%	ALL
Labeller L1	CMU	31.7/28.2	31.7/28.4	31.7/29.1	31.7/32.7
	+50%		28.2/28.4	28.2/29.1	28.2/32.7
	+75%			28.4/29.1	28.4/32.7
	+90%				29.1/32.7

Table 12 : Pairwise comparison scores in % according to the phonetic lexicon

4.4 Influence of the training of the phone models

With CoALT, we compared two automatic labellers using two ways to train the acoustic models on the Train corpus. For the labeller L1, the training stage used the sequence of phonemes of a sentence given by the manual labelling; the set contained 43 models. For L2, only the sequence of words was used and the phonetic transcriptions of every word were extracted from the CMU Pronouncing Dictionary v.0.6; the set contained 3 acoustic models less than for L1: /q/, /dx/ and /ax/.
Both automatic labellers used the phonetic lexicon including the pronunciation variants providing a coverage rate of 50%.
Using manual labelling for training models improves the automatic labelling score by 0.6% (from 28.8% to 28.2%). This improvement is significant but weak considering the costly and time consuming effort required by manual labelling. As the labeller L2 did not contain the model /ax/, we added the phonological rule [ax $\Rightarrow$ ah], as expected the gap between the two comparison scores reduces, (respectively 27.9% and 27.5%).

4.5 Summary

With CoALT, we have tested several parameters of our automatic labeller for English. We can conclude that increasing the number of pdfs of the acoustic models is not useful and even degrades the performance. On the other hand, adding common pronunciation variants improves the labelling performance. Moreover the manual labelling is not useful for training acoustic models but for extracting relevant pronunciation variants.

5. Conclusion

In this paper, we presented a tool that compares the results of two automatic labelling tools to a reference manual labelling in order to rank them. The ranking is based on the computation of insertions, deletions, substitutions, and shift between boundaries.
Besides ranking, the tool provides information about errors made by each automatic labelling tool and emphasizes their differences (for instance insertion, deletion).
The main feature of our software is that a user can define its own criteria for evaluating and comparing two automatic labelling tools.
Moreover, CoALT can be used for different languages provided that the user defines the phonetic alphabet and the optional rules. CoALT will soon be available under GPL.

6. References

Kim, S., Lee, K. and Chung, M. (2011). A Corpus-Based Study of English Pronunciation Variations. In *Proc. INTERSPEECH 2011*. Firenze, pp.1893--1996.

Dobrisek, S. and Mihelic, F. (2011). Time- and Acoustic-Mediated Alignment Algorithms for Speech Recognition Evaluation, In *Proceedings of*

INTERSPEECH 2001. Firenze, pp. 1517--1520.

Kawai, H. and Toda, T. (2004). An Evaluation of Automatic Phone Segmentation for Concatenative Speech Synthesis. In *Proceedings of ICASSP 2004*, Montreal, pp. 677--680,

Lee, K.-F. And Hon, H.-W. (1989). Speaker-Independent Phone Recognition Using Hidden Markov Models. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, v37 n11, pp 1641—1648.

Toledano, D. and Gomez, L. (2003). Automatic Phonetic Segmentation. *IEEE Trans. on Speech and Audio Processing*, v11, n6, pp. 617--625.

Wesenick, M.-B. and Kipp, A. (1996). Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals, In *Proceedings of ICSLP 1996*, Philadelphia, pp. 129--132.