Speech errors on frequently observed homophones in French: Perceptual evaluation *vs* automatic classification

Rena Nemoto, Ioana Vasilescu, Martine Adda-Decker
LIMSI-CNRS, BP 133, F91403Cedex France
E-mail: {rena.nemoto, ioana.vasilescu, madda}@limsi.fr

Abstract

The present contribution aims at increasing our understanding of automatic speech recognition (ASR) errors involving frequent homophone or almost homophone words by confronting them to perceptual results. The long-term aim is to improve acoustic modelling of these items to reduce automatic transcription errors. A first question of interest is whether homophone words such as *et*, (and) and *est* (to be), for which ASR systems rely on language model weights, can be discriminated in a perceptual transcription test with similar n-gram constraints. A second question concerns the acoustic separability of the two homophone words using appropriate acoustic and prosodic attributes. The perceptual test reveals that even though automatic and perceptual errors correlate positively, human listeners in conditions attempting to approximate the information available for decision for a 4-gram language model deal with local ambiguity more efficiently than ASR systems. The corresponding acoustic analysis shows that the homophone words may be distinguished thanks to relevant acoustic and prosodic attributes. A first experiment in automatic classification of the two words using data mining techniques highlights the role of the prosodic (duration and voicing) and contextual information (co-occurrence of pauses). Preliminary results suggests that additional levels of information may be considered in order to efficiently represent and factorize the word variants observed in speech and to improve the automatic speech transcription.

1. Introduction

The present contribution aims at increasing our understanding of automatic speech recognition (ASR) errors involving frequent homophone or almost homophone words by confronting them to perceptual results. The long-term aim is to improve acoustic modelling of these items to reduce automatic transcription errors.

During the last decade, several studies have established that human accuracy significantly outperforms machine accuracy on transcription tasks. These observations are particularly true when a large embedding context (complete and long sentences) is provided. They highlight that aspects of variation, such as pronunciation variants, noise, disfluencies, ungrammatical sentences, accents, which still remain important challenges for current automatic speech recognition systems, are well managed by human listeners. Word error rates of an order of magnitude higher were reported for ASR systems as compared to human listeners on English sentences taken from read continuous speech (CSR'94 spoke 10 and CSR'95 Hub3) databases under various SNR (signal-to-noise ratio) and microphone conditions (Deshmukh et al., 1996). A similar gap in performance between humans and automatic decoders has been reported for spontaneous speech (Lippmann, 1997). An interesting study (Shinozaki et al., 2003) in Japanese aimed at reproducing contextual information conditions of automatic speech decoders for human perception experiments. Stimuli comprising one target word embedded in a one word left/right context allow simulating word bigram networks as used by automatic decoders. In this very limited context condition, results indicate degraded human performances compared to the previous studies: error rate gap between humans an ASR systems no longer corresponds to an order of magnitude. Nonetheless they remain roughly half those of the recognizers. The comparison of these different studies highlights the importance of lexical context for accurate human transcription, the information is not exclusively locally grasped from the acoustic signal.

In line with (Shinozaki et al., 2003), this contribution aims at providing more insight on human speech transcription accuracy in conditions that reproduce those of state-of-the-art ASR systems, although in a much focused situation. We investigate a case study involving the most common errors encountered in automatic transcription of French: the confusion between, and more generally speaking, the erroneous transcription of two homophonic words *et* ("and") and *est* ("to be"). By focusing on this very particular case, we raise the question of whether humans use any quantifiable information for such homophone disambiguation that has not been exploited by ASR systems so far.

The second point investigated here following and complementing the perceptual study, concerns the acoustic separability of homophone words using appropriate acoustic and prosodic attributes. The frequency of the below studied items *et* and *est* can be related to their polisemy and propensity to occur in a large variety of contexts. However, the two words correspond to different part of speech, i.e. coordinative conjunction (et) and third person singular present-tense of the verb "to be" (est). Consequently, they occupy distinct positions within prosodic words and more largely, within sentences. These differences in terms of grammatical behaviour enable to believe the existence of acoustic and prosodic peculiarities of the two words which might possibly help humans to disambiguate them.

Finally, the proposed study also contributes to describe and compare factors of automatic vs perceptual confusability.

2. Corpus

We make use of the French *Technolangue*-ESTER corpus (Gravier and al., 2005), consisting in recordings of broadcast news shows from different francophone (French and Moroccan) radio stations. Transcription errors were extracted from the automatic transcriptions produced by the LIMSI speech recognition system developed for the 2005 ESTER evaluation (Gauvain et al., 2005). This system achieved the best results with an overall performance of 11.9 WER for the speech transcription task (Galliano et al., 2005). The ASR system made use of 4-gram language models (LM) and context-dependent acoustic phone models. Among the frequent words, *et* and *est* are the most error-prone items: 25% of *et* "and" and 20% of *est* "is" (verb "to be") occurrences are misrecognized (Adda-Decker, 2006).

3. Automatic transcription errors

Several reasons may be enumerated for ASR errors: OOV (out of vocabulary) words, words and word sequences which seldom occur in training data, acoustic confusability due to homophones. The French language is particularly challenging for automatic transcription: it admits a large number of homophones (especially different verb forms, e.g. *tuer, tué, tués...* "to kill"), almost all phonemes (both vowels and consonants) correspond to monophonic function words, and many of them admit homophones, for example à, "at", "to", *a*, and *as*, present tense conjugated forms of "to have"; *et*, "and" and *est*, "to be", etc. Such words are particularly frequent and often less carefully pronounced (i.e. hypo-articulated). Whereas the overall WER of the LIMSI system is below 12%, error rates from the 20 most frequent words contribute to more than one fourth to these transcription errors. We focus here on the two most frequent homophone words *et* and *est*. Although the canonical pronunciation of *est* corresponds to a mid-open vowel [ɛ], in fluent speech its actual realization tends to become a closed [e], homophone with the pronunciation of *et*. Section 4 below is dedicated to the perceptual evaluation of different error types involving the two homophone words. The aim of the experiment is to check if human transcribers are able to correctly identify the two words in limited contexts which have proved to be ambiguous for the ASR system. In a second part (Sections 5 and 6), the acoustic separability of the two words is explored, using appropriate acoustic and prosodic attributes. Integration of perceptual findings and acoustic measurements is discussed in section 7.

4. Perceptual evaluation

The perceptual experimentation on the automatic transcription errors of the *et*/*est* homophones has been conducted with the aim of clarifying whether human word perception outperforms automatic decoding of the target words in a 7-gram (i.e. 4-gram left and 4-gram right) word context. Table 1 below, shows some typical examples of transcription errors involving the target words *et* and *est*. The excerpts shown contain the target word in the middle of a 7-gram and are surrounded by 3 left and right neighbouring words, thus integrating the maximum scope of the language model for the target word transcription. In many situation however, the ASR system backs off for lower n-grams, resulting in less than 7 words.

In particular, two questions have been addressed: (1) are the human transcriptions on the homophones *et/est* more accurate than the automatic ones in conditions corresponding to contextual N-gram constraints similar to those of automatic speech decoding; (2) if humans are more competitive, which of the linguistic levels of information (syntactic, semantic, prosodic, voice quality …) may have potentially contributed.

Ex.1	
REF HYP	rhume de cerveau **est** la maladie virale rhume de cerveau **et** la maladie virale
Ex.2	
REF HYP	sur les salaries **est si** formidable que sur les salaries **ici** formidable que
Ex.3	
REF HYP	politique aujourd'hui il **est** essential d'approfondir politique aujourd'hui il essential d'approfondir

Table 1: Examples of 7-gram stimuli with different types of errors: *et/est* confusion (ex.1 "cold fever is the viral disease"), *est* within a syntagm substituted by another word (Ex. 2 "on the salaries is so formidable that"), *est* deletion (Ex. 3 "politics today it is essential to go into detail").

4.1 Test material selection

Stimuli comprising the target *et/est* homophones in limited n-gram contexts are selected. The test material consisted in 83 chunks extracted from the ESTER development corpus (dev04). We call chunk a 7-word string with the target word as center (Table 1). Forced alignment of the reference manual transcriptions is carried out and selected chunks are extracted automatically.

The choice of 7-gram chunks aims as providing the human subjects as much information around the target word as used by a 4-gram LM-based transcription system in optimal conditions. Stimuli mainly contain an erroneously transcribed *et* or *est* in central position (68 stimuli). They also illustrate different types of errors observed in the ESTER development corpus: insertions, deletions, substitutions of the target words only or of the target words together with surrounding words (target word within a syntagm). Selected errors aim at covering all the erroneous transcription case figures encountered in the ESTER dev04 corpus (as illustrated in Table 1, Ex. 1, 2, 3 above). Some distracting items consisting in 7-gram chunks correctly transcribed as well as different target words were also added (see Table 2).

Ex.1	
REF	recréer un intérieur bourgeois le décor ne
HYP	recréer un intérieur bourgeois le décor ne
Ex.2	
REF	tristesse inouïe c'**est** un ogre c'
HYP	tristesse inouïe c'**est** un ogre c'

Table 2: Examples of 7-grams distracting stimulus (Ex. "recreate a bourgeois interior, the decorating don't") and correctly transcribed stimulus (Ex. 2 "incredible sadness, this is a monster this").

Table 3 below sums up the different types of chunks corresponding to contexts giving rise or not to automatic transcription errors.

Chunks (nbr.)	**Types of errors**
5 distractors	Stimuli without *et/est* in the middle
10 corrects	Stimuli with *et/est* correctly transcribed by the system
20 *et/est* symmetric confusions	Stimuli with symmetric ASR confusions of *et/est*
48 other errors (6/type/target word)	Stimuli with other errors: insertions, deletions, erroneous transcription of target word alone or within a syntagm.

Table 3: Types of automatic transcription errors illustrated by the 83 selected stimuli.

4.2 Test protocol

Sixty native French subjects took part in the experiment. They were not informed of either the target words or the selection criteria, or the fixed chunk length.

The 60 subjects were divided into two sub-groups according to different test conditions as following:

40 subjects underwent an **acoustic+language model (AM+LM)** condition test. The 83 stimuli have been submitted to two groups of 20 subjects *via* a web available interface. Listeners were provided with the audio excerpt corresponding to the 7-gram chunk and had to transcribe the entire chunk. Each group of 20 subjects listened to and transcribed half of the stimuli. This choice was made to limit the duration of perceptual test to less than one hour: subjects were spending about 20 x *RT* (real-time) to transcribe a stimulus (compared to automatic transcriptions processed in 10 x *RT*). The two groups were comparable in terms of age and background.

20 subjects performed a local **language model (LM)** condition test on the 30 chunks focusing on *et* vs *est* confusion (i.e. the stimuli for which the system transcribed *et* by *est* and vice-versa to which we added the 10 correct chunks as control stimuli). They had to fill in the written version of the 30 chunks using the most plausible item *et* or *est*, as suggested by the 3-word left and right contexts. Figure 1 below gives schematic representation of the written test protocol. This condition is a simplification of the ASR ambiguity processing, which has to score all possible expanded ambiguities of the uttered sequence. The rationale of this test is twofold: syntactic/semantic information of the written sequence contributes to solve ambiguity; humans explicitly focus on local ambiguity. This test assumes perfect homophony for the target.

et
Rhume de cerveau la maladie virale
est

Figure 1: Written test corresponding to a local LM condition.

4.3 Results

Results are measured in terms of erroneous transcription of the target words compared to the reference transcriptions. Human error rates are then compared to ASR word error rates.

As a general observation, subjects produced an average of 6 errors per person on the chunk-central words independently of the test conditions. Detailed WER rates for the different stimuli sets and conditions are reported in Table 4. We have to recall that the ASR error was the criterion for stimuli selection.

	WER (word error rates)		
Stimuli	**ASR**	**Humans**	
	AM+LM	**AM+LM**	**LM**
5 distractors	0	0	-
10 corrects	0	1.4	8.2
20 *et/est* symmetric confusions	100	25.5	27.6
48 other errors (6/type/target word)	100	16.0	-

Table 4: WER on 4 stimuli subsets in different automatic/human transcription conditions: ASR (selection criteria); LM (written test on local ambiguity); AM+LM (audio test).

For the **AM+LM condition**, results of the perceptual test show that no error are produced on the distractor stimuli and that a marginal error rate (1.4%) is measured on the 10 perfectly decoded stimuli by the ASR system. However on the stimuli subset corresponding to system confusions, an important increase in the human error rate can be observed. A statistical significance test was carried out to measure the validity of this result. The potential correlation between human and automatic transcription solutions has been checked statistically (with one factor "system answer for target word'" ANOVA, using "correct" vs. "erroneous" as nominals). The factor "system answer for the target word" is statistically significant for both LM (F(34,07), p<0.0001) and AM+LM (F(38,22), p<0.0001) conditions.

Consequently, human produce more errors on stimuli misrecognized by the ASR system. Reversely humans are almost error free on the correctly decoded stimuli. When trying to weight the achieved perceptual results in order to reflect the distribution of the different stimuli types in the news speech corpora, humans appear to be 4-5 times more accurate than ASR system in this particular test condition.

We also checked which of the two words *et* vs *est* is more ambiguous. An ANOVA analysis (with one factor "target word", using "*et*" and "*est*" as nominals) showed that *est* have been missed more frequently by human listeners than *et* (F(38,95), p<0.001).

Finally, when looking at different types of errors for each of the target words, namely insertions, deletions and erroneous transcriptions of the target word or of the target word and the surrounding words, one may notice that the type of error and the number of errors produced by the listeners are positively correlated: the more ambiguous the local context the more frequently the correct solution is missed. Consequently, humans produce more frequently errors for the stimuli for which the system missed the target word and the surrounding context than for the stimuli for which the target word has been only deleted or inserted while the other surrounding words remaining correctly transcribed.

This finding suggests that the local linguistic ambiguity is problematic for both the ASR system and the humans. In case of local ambiguity the transcription forces "random" choices which are prone to error both for humans and ASR systems.

The **LM test** might be considered as the easiest one, as only a local ambiguity has to be worked out, while relying on the surrounding written words. We remind that the LM test represents the written version of the stimuli focusing on the symmetric *et/est* confusion. However the lack of punctuation (due to the ASR simulation protocol) probably adds some difficulty here. We compared the results for the LM test with the LM+AM test section focusing uniquely on the symmetric *et/est* confusion. The difference between the two conditions is statistically significant (one sample t-test, p<0.0025, t=2.66, p=0.0078) and the LM condition generates more errors than AM+LM condition. Better results on the AM+LM condition might be related to additional structuring information of the audio signal: the lack of punctuation does most likely not allow retrieving information on syntactic structures which might rely on prosodic cues in AM+LM condition. However, no statistical difference has been observed when comparing the ratings for each of the target words, i.e. subjects are equally competitive in processing chunks with *et* or *est* and the target words seem to be equally ambiguous in the given word strings. This information suggests that the two polysemic words spawn comparable contexts in terms of intrinsic degree of ambiguity in the French language and human subjects encounter similar challenges in processing them. It suggests that for the given string length, both humans and ASR systems leave some unresolved ambiguities, even though less numerous in the case of humans (at least as observed in this perceptual experimentation).

4.4 Discussion on perceptual evaluation

When comparing ASR vs human results, we observe that almost no errors occur for the distractor stimuli and for the 10 stimuli without confusions on the target words. The mean error rate per person is 6, even though error rate varies strongly with the type of local context. The context entailing symmetric *et/est* errors for ASR are thus highly ambiguous as well as contexts for which the local ambiguity concerns the target homophone word and the close surrounding context. Among the two homophone words, *est* is more frequently misrecognized by the human listeners than *et* (25% vs 10%). A comparison between the system answers and the human transcriptions, reveals that humans achieve better results in terms of correct *et/est* ratings for those stimuli correctly transcribed by the ASR system as well.

The perceptual test reveals that even though automatic and perceptual errors correlate positively, in conditions which attempt to approximate the information available for decision for a 4-gram language model, human listeners deal with local ambiguity more efficiently than the ASR system. Perceptual results seem to support the following hypothesis: differences in ratings for similar ambiguous syntactic structures suggest prosodic/acoustic information may help in operating the right choice in terms of target word selection.

5. Automatic classification with data mining techniques

Perceptual evaluation suggested that humans are globally 5 times better than the automatic system in conditions that approximate the amount of information available for the latter. However, we are aware of the fact that even if we try to stick as close as possible to ASR system conditions, the comparison remains "unfair". Humans have access to local syntactic and semantic information combined with a more extended knowledge of the broadcast news background. Besides, prosodic and voice quality information may interplay with the lexical information.

We aimed here at a more in-depth investigation of the **acoustic and prosodic information** which potentially contributed to the correct classification of the homophone words *et* and *est* and which have been neglected by the classical acoustic parameters (i.e. cepstral vectors) and the acoustic models (HMMs). We made use of data mining techniques in order to automatically classify the homophone words *et/est* thanks to a collection of acoustic and prosodic attributes.

5.1 Corpus and extraction methodology

Homophone words *et* and *est* have been extracted (*cf.* Table 5) from 55 hours of speech coming from different French broadcast news (BN) channels (France Inter, Radio France International, France Info, Radio-television of Morocco) from the *Technolangue*-ESTER corpus. Several acoustic and prosodic parameters have been defined and automatically extracted thanks to the Praat software (Boersma and Weenink, 1999) and to the LIMSI automatic speech alignment system (Gauvain et al., 2005). Selected parameters concern duration, fundamental frequency, formants (F1, F2, and F3) and surrounding context (pauses preceding/following the target word).

For pitch and formant values, measures have been carried out on a frame by frame basis every 5*ms*. For each segment, a voicing ratio was computed as the ratio between the number of voiced frames and the total number of frames. For each segment, mean values have been computed for the parameters f0 and formants over all voiced frames of the segment.

Words	Occurrences	Phonemes
et	19.1k /e/	[e]
est	14.5k /ɛ/	[ɛ] 5.0k, [e] 9.5k

Table 5: Occurrences of *et* and *est* in the BN corpus.

5.2 Acoustic analysis

Prior to the automatic classification, we focused on the acoustic and prosodic parameters potentially able to differentiate the homophone word pair. Duration, voicing characteristics and pauses before and after the homophone words have been considered.

5.2.1. Duration

Figure 2 below represents *et* and *est* duration distribution. Durations range from 30 to 200 ms for both words *et* and *est*.

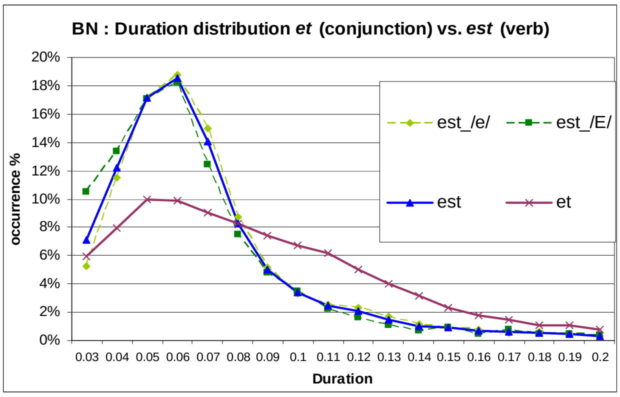

Figure 2: Duration distribution of the homophone words *et/est*: "*et* (in red)" and "*est* (in blue)" (/e/ in clear green and /ɛ/ in dark green). Different lines correspond to number (in %) of occurrences per duration threshold.

Duration's distribution comparison for the homophone words *et/est* shows differences between the two target words. The conjunction *et* has a relatively flat distribution, including in particular more segments with durations above 80 ms, whereas *est* has an almost bell-shaped distribution centered on 60 ms. On average, the conjunction *et* lasts longer than the verb *est*.

5.2.2. Fundamental frequency (f0)

As mentioned in the introduction, the two homophone words correspond to different parts of speech (conjunction *vs* verb). This distinction may be linked to the prosodic realization of words, e.g. the duration of the words and the fundamental frequency (f0). One may hypothesize that a verb inside a prosodic word is differently realized in terms of average f0 from a conjunction occurring at the beginning of a prosodic word and serving in isolating syntactic blocs. Furthermore, the voicing may vary according to the position of a lexical item within a prosodic word, e.g. the voicing may be partial at the beginning of the prosodic word in particular when the prosodic word is preceded by caesuras or pauses.

We address here the question of the function of the voicing ratio in the articulation of a word and we measure this parameter for the two homophone words *et* and *est*. The voicing ratio is computed as described above (section 5. 1) and corresponds to the percentage of non null f0 values. To analyze the extracted voicing ratio measures, three classes have been defined:

1. **Devoiced**: % of voicing from 0 to 20%;
2. **Partial voicing**: % of voicing from 20 to 80%;
3. **Voicing**: % of voicing from 80% to 100%.

For the three voicing ratio classes, Figure 3 shows the proportion of segments in each class. For each class, results are given first for *et*, next for *est*. For the latter two bars are added to separate [ɛ] from [e] pronunciation. To produce comparable results for the different conditions, absolute counts are transformed in relative rates which sum up to 100% to each condition. For both words, as expected, "devoiced" class contains a small amount of data. In the "partial voicing" class *et* is better represented that *est*, conversely, *est* is more frequent in the "voicing" category. The results suggest that the conjunction is less voiced than the verb[1].

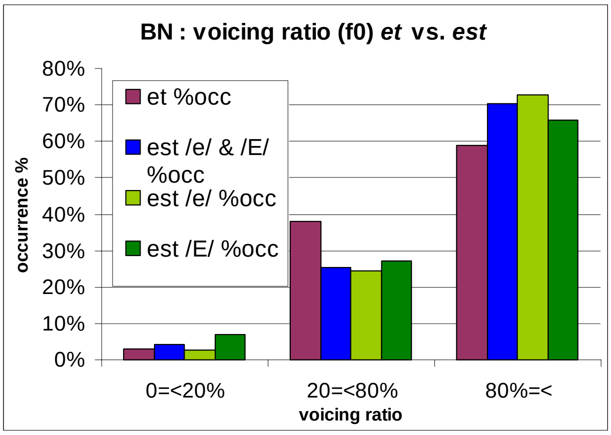

Figure 3: Histograms of the two homophone word occurrence distributions according to the voicing ratio: "*et* (in red)" and "*est* (in blue)" (/e/ in clear green and /ɛ/ in dark green).

5.2.3. Left-right pause co-occurrences

The pauses play an important role in the process of automatic prosodic information extraction (Lacheret-Dujour, Beaugendre, 1999). We aimed here evaluating the relation between the pause (we include here in the class "pause" silences, breaths and filled pauses, i.e. hesitations) and the analyzed homophone words. We thus

[1] A similar analysis on *a* (the auxiliary verb « avoir » to have) and *à* (the preposition « to ») shows common voicing patterns allowing generalizing these observations.

examine their left-right co-occurrences with the target words *et* and *est*.

Table 6 presents the percentage of occurrences of left and right pauses. The main difference between the conjunction (*et*) and the verb (*est*) concerns the amount of pause occurrences, in particular left pauses. One may hypothesize that the verbs (*est*) are less frequently preceded by a pause than the conjunctions (*et*).

Words	***et***	***est***
Left pause	**49%**	9%
Right pause	7%	5%

Table 6: Left and right pause occurrences (in %) for the target words *et/est* (silence, breath, hesitation)

6. Attribute definition

In section 5 three parameters have been analyzed in order to model the prosodic characteristics of the homophone words *est* and *et*. The analysis of duration, voicing ratio and presence of pauses before/after the target words highlighted some differences between the two words. This preliminary result encouraged us in searching for acoustic and prosodic parameters to model differences between the two words. In this section we address the matter of the automatic separability of the two words thanks to appropriate acoustic and prosodic attributes (Nemoto et al., 2008).

41 acoustic and prosodic attributes have been selected for the automatic classification. They were chosen in order to model both the target word (intra-phonemic attributes) and its relation to the context (inter-phonemic attributes).

Intra-phonemic attributes (33): duration, f0, voicing ratio, first three formants (global mean values by segments and begin, center, end values). We also calculated the differences (Δ) between begin-center, center-end and begin-end for the f0 and the formants.

Inter-phonemic attributes (8): duration, f0, pauses. Duration attributes were measured as following: the difference between a center segment duration of target word and a center segment duration of a previous/following vowel, even though there are consonants or pauses between these phonemes. For the f0 and the formants, Δ values were calculated as the difference between the mean values of the target word vowel and the previous/following vowel. In addition, the difference between two mean values of previous and following vowels of the target word was considered as well. Finally, left-right pause attributes were added too.

6.1 Classification experiment

To automatically classify the homophone words with the 41 attributes defined above, we tested 25 algorithms (Bayesian classifiers, Trees, SVM, etc.) implemented in the data mining software Weka (Witten & Frank, 2005). The classification experiments were performed using a cross-validation method. Table 7 gives the results of correct word identification by automatic classification. Identification scores are classified according to the best algorithm (LMT), the mean of 10 best algorithms and as a mean of all 25 tested algorithms.

6.2 Attribute selection

41 attributes were selected to classify the homophone words *et* and *est*. We hypothesize that among the 41 attributes some are more discriminatory than others. 10 best attributes have then been selected thanks to the LMT (Logistic Model Trees) algorithm which provided the highest result. As for the 41 attributes, the percentage of correct word identification has been computed with the 10 best attributes. The selected attributes are listed in Table 8 and their results are presented in Table 7. The results with 10 best attributes are slightly less efficient than with 41, but the difference remains low. This result suggests that only 10 attributes are almost as discriminatory as 41. The question is then to identify the most efficient features encoded in 10 attributes.

	et* vs. *est	
Treated attribute numbers	10	41
Best algorithm (LMT)	77.0	78.0
Mean of best 10 tested algorithms	75.7	76.1
Mean of all 25 tested algorithms	70.2	69.3

Table 7: Comparison of correct word identification % according to number of attributes and algorithms.

Words	***et* vs. *est***
1	**Left pause**
2	**Δ Left duration**
3	*ΔF1 begin - center*
4	**Δ Right duration**
5	**Δf0 Right**
6	*duration*
7	**Δf0 Left**
8	*f0 voicing ratio*
9	**Δf0 Left-Right**
10	**Δ L-R duration**

Table 8: 10 better classified attributes by the LMT algorithm (intra-phonemic attributes in italic type and inter-phonemic ones in bold type).

Table 6 lists the most efficient attributes as produced by the LMT classification. In line with the acoustic measures, they are related to duration parameters (phoneme duration and Δ inter-phonemic duration), f0 (Δ inter-phonemic f0, voicing) and pause preceding the target word. These results suggest that prosodic features are relevant in distinguishing the two words. Among the phonetic features, intra-phonemic F2 values are particularly salient.

6.3 Discussion on automatic classification

In this section, different acoustic realizations of the two frequent homophone words *et* and *est* have been examined. The comparison of durations distribution showed that the word *et* tends to last more than the verb *est*. This simple measure suggests that homophones, realized a priori with the same phonemes (for example, the same formant values for the vowels), may differ in their prosodic realization. In addition, differences in voicing have been also noticed.

Next, we defined 41 intra- and inter- phonemic acoustic and prosodic measures potentially relevant for the automatic classification of the two words and we tested different algorithms implemented in the Weka software. Results are promising: classification scores are around 70% of correct identification for *et* vs *est*. The automatic classification results illustrate that the attributes concerning intra- and inter-segmental duration, as well as voicing and differences in f0 between the target segment and close context, are particularly robust.

7. Conclusion

The present contribution aimed at increasing our understanding of automatic speech recognition (ASR) errors involving frequent homophone or almost homophone words by confronting them to perceptual results. The longer-term aim is to improve acoustic modelling of these items to reduce automatic transcription errors.

A first question of interest addressed in this paper is whether homophone words such as *et*, “and” and *est*, “to be”, for which ASR systems rely on language model weights, can be discriminated in a perceptual transcription test with similar n-gram constraints. A second question concerns the acoustic separability of the two homophone words using appropriate acoustic and prosodic attributes.

A perceptual test has been conducted in order to evaluate human subjects’ capacity to correctly transcribe the two homophone words in ambiguous contexts. Perceptual results have been measured in terms of erroneous transcription of the target words compared to the reference transcriptions. Human error rates were then compared to ASR word error rates. Human transcriptions’ analysis showed that distractor stimuli were error-free. A marginal error rate has been measured on the perfectly decoded stimuli by the ASR system. Reversely, on the stimuli subset corresponding to system confusions, an important increase in the human error rate could also be observed. Results suggest that local contextual ambiguity is problematic for both the ASR system and the humans.

The corresponding acoustic analysis shows that the two homophone words *et* “and” and *est* “to be” may be distinguished thanks to some relevant acoustic and prosodic attributes. A first experiment in automatic classification of the two words using data mining techniques highlights the role of the prosodic (duration and voicing) and contextual information (co-occurrence of pauses) in distinguishing the target words.

8. Acknowledgements

This work was partially financed by the research cluster Digiteo through a Région Ile-de-France doctoral grant for the first author and by OSEO under the Quaero program.

9. References

Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l’analyse linguistique de corpus oraux, In Proceedings of *JEP*.

Boersma, P., Weenink, D. (1999-2007). Praat: doing phonetics by computer, www.praat.org.

Deshmuk, N., Dunca, R.J., Ganapathiraju, A., Picone, J. (1996). Benchmarking human performance for continous speech recognition, in Proc. of *ICSLP*.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news, In Proceedings of *EUROSPEECH*, Lisbonne.

Gauvain, J.L., Adda, G., Adda-Decker M., Allauzen A., Gendner V., Lamel, L., Schwenk, H. (2005). Where Are We in Transcribing French Broadcast News? In Proceedings of *Interspeech*, Lisbon.

Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., Mc Tait, K., Choukri, K. (2004). ESTER, une campagne d’´evaluation des systèmes d’indexation automatique d’émissions radiophoniques en français, In Proceedings of *JEP-TALN*.

Lacheret-Dujour, A., Beaugendre, F. (1999). *La prosodie du français*. CNRS, Paris.

Lippmann, N. (2003). Speech recognition by machines and humans: benchmarking human performance for continous speech recognition, In *Speech communication*, 22(99): 1-15.

Nemoto, R., Adda-Decker, M., Vasilescu, I. (2008). Fouille de données audio pour la classification automatique de mots homophones. In *EGC’2008*, Sophia-Antipolis, France.

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research, In *Speech Communication*, 49(5): 336-347.

Selkir, E. (1996). The prosodic structure of function words. In *Signal to syntax: bootstrapping from speech to grammar in early acquisition*. Mahwah, Lawrence Erlbaum, pp 187-214.

Shinozaki, T., Furui, S. (2003). An assessment of automatic recognition techniques for spontaneous speech in comparison with human performance. Proceedings of *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Witten, I. H., Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco.