INTERSPEECH 2009 BRIGHTON

A perceptual investigation of speech transcription errors involving frequent near-homophones in French and American English

Ioana Vasilescu.[1], *Martine Adda-Decker.*[1], *Lori Lamel.*[1], *Pierre Hallé*[2]

[1] Spoken Language Processing Group, LIMSI-CNRS, 91403 Orsay, France

[2]LPP-CNRS, 19 rue des Bernardins, 75005 Paris, France

ioana@limsi.fr, madda@limsi.fr, lamel@limsi.fr, pierre.halle@univ-paris5.fr

Abstract

This article compares the errors made by automatic speech recognizers to those made by humans for near-homophones in American English and French. This exploratory study focuses on the impact of limited word context and the potential resulting ambiguities for automatic speech recognition (ASR) systems and human listeners. Perceptual experiments using 7-gram chunks centered on incorrect or correct words output by an ASR system, show that humans make significantly more transcription errors on the first type of stimuli, thus highlighting the local ambiguity. The long-term aim of this study is to improve the modeling of such ambiguous items in order to reduce ASR errors.

Index Terms: American English, French, ASR, speech perception, speech ambiguity, near-homophones

1. Introduction

During the last decade, several studies have established that humans significantly outperform machines on speech transcription tasks. These observations are particularly true when large surrounding contexts (complete and long sentences) are available. The studies demonstrated that human listeners are better at handling many aspects of variation, such as pronunciation variants, noise, disfluencies, ungrammatical sentences, accents, which are still important challenges for current ASR systems.

The word error rates reported for automatic speech recognition (ASR) systems were an order of magnitude higher than those of human listeners on English sentences from read continuous speech (CSR'94 spoke 10 and CSR'95 Hub3) databases under various SNR (signal-to-noise ratio) and microphone conditions [2]. A similar difference in performance between humans and automatic decoders has been reported for spontaneous speech [3]. An interesting study [4] on Japanese aimed at reproducing the contextual information conditions of automatic speech decoders in human perception experiments. Stimuli comprising one target word embedded in a one word left/right context allow the simulation of word bigram networks as used by automatic decoders. In this very limited context condition, results indicated degraded human performance in comparison to the previous studies [2, 3]. The different in performance between humans and machines is reduced to about a factor of 2 from an order of magnitude. The comparison of these studies highlights the importance of lexical context for accurate human transcription, the information is not taken exclusively from the local acoustic region.

Similar to the study reported in [4], this contribution aims at providing more insight on human speech transcription accuracy under conditions simulating those of state-of-the-art ASR systems, in a very focused situation. We investigate a case study involving the most common errors encountered in automatic transcription of American English and French: the confusion between, and more generally speaking the erroneous transcription of, two homophonic words, **and** and **in** in English, and **et** ("and") and **est** ("is") in French. The word error rates (WER) for the English words are is about 15%, whereas the French homophonic word error rates are over 20% in broadcast news data [1]. By focusing on this very particular case, we raise the question of the information potentially exploited by human listeners and ignored by ASR systems to disambiguate such homophonic words. The long-term aim of this work is to improve the modeling of such ambiguous items to reduce ASR errors.

The next section is dedicated to the working hypothesis underlying the perceptual experiments. The speech corpora in American English and French are described in Section 3 and related automatic speech recognition errors in Section 4. Section 5 details the experimental setup with the results of the perceptual experiments discussed in Section 6, before concluding in Section 7.

2. Working hypothesis

In the current study we presented limited length audio stimuli to human subjects to test their transcription capacities on the central word. The stimuli include as much context as optimally used for a word decision by ASR systems. A comparison of human transcriptions of the central word with those of ASR systems, may then be indicative of either the **intrinsic ambiguity** of the stimuli in the case of joint human and ASR errors, or of ASR limitations due to simplified modeling hypotheses. We refer to the latter as the **model bias**. ASR transcription errors can then be viewed as arising from "ambiguous speech regions", which are due either to intrinsic ambiguity of the speech signal or to the model bias. The intrinsic ambiguity hypothesis concerns the case where both ASR and humans produce errors on the target words. In this particular case both the central word acoustics and the local context provided by the neighboring words remain ambiguous, and neither human perception nor the ASR system could solve this ambiguity. This case clearly requires a larger context to potentially solve ambiguity. Further work is needed to estimate the optimal additional information required for both human perception and ASR systems. The hypothesis of the model bias may be supported by the stimuli carrying an ASR error, but correctly transcribed by the human subjects. Here some information used by humans is lacking in the model. If the central word and its erroneous counterpart are homophones or near-homophones, we hypothesize that the information for the right decision should mainly come from the

surrounding words. Hence the language model (LM) has the major responsibility in the erroneous decision here. For non homophone words, both the LM and the acoustic model (AM) may contribute to the wrong decision. For the present study we propose the following loose definition of **near-homophones**: they are minimal pairs such as *have, had* in American English or *été* (/ete/, "been"), *était* (/etɛ/, "was") in French, where words differ by only one acoustically close phoneme. Near-homophones may also arise from reduced pronunciations, where a word such as *and* (/ænd/, reductions [æn] or even [n̩]) may become near-homophones or homophones with *in* (/ɪn/, reduction [n̩]). In order to limit the risk of drawing conclusions which are too language-specific, the study is conducted on two languages, French and English. The main focus however concerns the question of whether erroneously ASR-transcribed stimuli also prove to be harder for human perception or whether humans rely on some crucial information missing in ASR speech models.

3. Corpus

For American English, the study made use of a subset of the NIST HUB4 corpus of broadcast news shows from different radio stations (VOA, ABC, etc.). The stimuli were selected from the EARS dev03 development data (dev03nist). They are comprised of about 2.5 hours corresponding to 24.7k words. The overall word error rate of the hypothesis used in this study data is 11.2%. In French, data from the TECHNOLANGUE-ESTER corpus [6], which consists of Francophone (French and Moroccan) broadcast news was used. The corpus was designed for the national TECHNOLANGUE campaign (evaluation of language technologies for the French language) ESTER focusing on rich transcription and indexing of radio broadcast news in French. The subset of data considered here corresponds to the ESTER development (dev04) corpus with 10 hours of broadcast speech corresponding to 94k transcribed word tokens. The word error rate of this data was about 12% [5]. For both languages most of the errors (over 65%) are word substitutions, with twice as many deletions as insertions.

4. Automatic speech transcription errors

Several reasons help explain the ASR errors: words not included in the system's lexicon, words and word sequences which seldom occur in the training data, acoustic confusability due to homophone or almost homophone words. Both languages contain a large number of such words, in particular monosyllabic function words, for example, *has/had/have, as/is, are/were, etc.* in American English; *et "and", est "is", à "to", a/as "has/have,singular present tense", il "he", y "there"* etc. in French. These words, particularly frequent, tend to be well predicted by context, however often hypoarticulated. They generally entail a large number of automatic transcription errors. For instance, the twenty most frequent words, as measured from huge written and transcribed audio corpora, are all monosyllabic function words and are involved in more than 25% of measured ASR errors, both in French and American English. For this work we have chosen among the most frequent erroneously transcribed near-homophone pairs, the pair (*et "and" vs est "is"*) in French and the pair (*and vs in*) in American English, each pair accounting for 5% of errors in both languages. In French, although the canonical pronunciation for *est* corresponds to a mid-open vowel /ɛ/, in fluent speech its actual realization tends to become a closed [e], a homophone with the pronunciation of *et* [9]. In American English, the reduced pronunciation of *and*, with a deleted word-final /d/, is a near-homophone of *in*, as both words have acoustically close vowel nuclei.

	Erroneous 7-gram excerpts
1	REF: of the day **and** it is almost HYP: of the day **in** it is almost
2	REF: politique aujourd'hui **est** essentiel d'approfondir HYP: politique aujourd'hui *** essentiel d'approfondir
3	REF: escape on tape *** the two were in HYP: escape on tape **and** the two were in
4	REF: de mai difficile **et** les syndicats HYP: de mai difficile **mais** les syndicats
5	REF: diverse social fabric **in Salt** Lake City HYP: diverse social fabrics **console** Lake City

Table 1: Examples of 7-words speech chunks (REF: reference transcription; HYP: ASR hypothesis) with different error types: (1) near-homophone substitution, (2) deletion, (3) insertion, (4) other word substitution, (5) multi-word substitution (syntagm), in American English and French (*politics today it is essential to go into detail*; *of May difficult and the trade unions*).

5. Experimental setup

The major controlling parameter for stimulus selection for this study of the link between ASR and human transcription errors was the presence of a word which has a high contribution to the ASR word error rate. This control then entails highly frequent words (with frequency ranks < 10 for both languages), which are monosyllabic and near-homophones (cf section 1), resulting in the choice of *et/est* and *and/in* as target words. As the LIMSI ASR system makes use of 4-gram language models (LM) [5], speech chunks of 7 words, corresponding to two 4-grams overlapping by the central word, were extracted for the perceptual experiment. The central word of these stimuli correspond to ASR errors involving at least one word of the selected word pairs in American English *and/in* and French *et/est*. The chunk length choice aims at providing the human subjects with as much information around the target word as maximally used by a 4-gram LM-based transcription system. In many situations however, the ASR system backs off for lower n-grams, resulting in a smaller than 7 word segment. Subjects were asked to transcribe the 7-gram chunks. Their performances were assessed for the central (4th) target word with respect to the target word type (correct/erroneous) of the ASR transcription. Table 1 shows some typical examples of 7-gram chunks in American English and French: the reference transcription with the central target word being either *and/in* or *et/est*, and the ASR hypothesis with different error types for the target word.

5.1. Error types and stimuli selection

The 7-gram speech chunks have been selected to cover various situations, including different error types as well as correctly transcribed chunks. The error types include substitutions, deletions and insertions in proportions similar to those of the ASR systems (majority of single word substitutions, few deletions and insertions). Most selected chunks involve a single word error, however a small proportion of chunks are composed of erroneous speech regions spanning several words around the central

Training/distracting and **error-free** 7-gram excerpts	
1	REF: airlines face the **possibility** of a strike HYP: airlines face the **possibility** of a strike
2	REF: réaliste le plateau **est** une coquille vide HYP: réaliste le plateau **est** une coquille vide

Table 2: Examples of (1) training/distracting and (2) error-free 7-word chunks in American English and French (*realistic the stage is an empty shell*.

target. Five error types, illustrated in Table 1, were identified as covering the majority of error situations produced by ASR systems: (1) substitutions between selected near-homophone word pairs (*and* by *in* and vice-versa;*et* by *est* and vice-versa); (2) omission of the target word; (3) insertion of the target word; (4) substitution of the target word by a word other than its near-homophone; and (5) multi-word substitution of the target word and possibly surrounding words: typically a syntagm involving the target word entirely transcribed by another syntagm. Finally, besides stimuli for the various ASR error types of the near-homophone pairs, some stimuli with the target words correctly transcribed by the ASR system, as well as *training* and *distracting* stimuli (with other words than the target ones as chunk centers) have been included (see Table 2). A prior forced alignment of the manual reference transcription had been carried out, the chunks were extracted automatically, and selected manually to fit the different error type proportions. In both languages the near-homophone substitutions are frequent and 20 such confusions were selected for each language. For the other less frequent error types, the exact number of extracted stimuli varies from French to English, resulting in different total numbers of selected stimuli for both languages. The selected chunks last about 1.5-2 seconds each, and were pronounced by different speakers, male and female.

• **American English experiment**. The stimuli set is comprised of 129 chunks extracted from the NIST HUB4 development corpus (dev03). Stimuli were selected to cover the typology of ASR errors as described above. 102 stimuli contain mainly erroneously transcribed and some more error-free *and/in* central word speech chunks. The remaining stimuli are training and distracting 7-grams (i.e. stimuli with another word than one of the target pair of words as 4th word of the speech chunk).

• **French experiment**. The test consists of 83 chunks extracted from the ESTER development corpus (dev04). They illustrate the main error types encountered in ESTER dev04 corpus. 78 stimuli contain erroneously transcribed and some additional error-free *et* or *est* in their central position. The 5 remaining stimuli are training and distracting stimuli.

5.2. Test protocol

21 native English (for the American English experiment) and 20 native French (for the French experiment) subjects provided transcriptions *via* a web-based interface. They were instructed that the aim of the experiment was to compare their manual transcription with the automatic transcription, however they were not informed of the target words nor of the selection criterion of sentences, nor of the fixed chunk length. Subjects were provided with the audio of the 7-gram chunks and were given very brief instructions to help them transcribe it. The stimuli were presented in a random order. Subjects could listen to each stimulus as often as they wished.

6. Perceptual results

In the following, human performance is measured on the central target words: *and/in* in American English and *et/est* in French. The human performance is assessed on the subsets of correctly and erroneously ASR-transcribed stimuli. Finally we investigate the potential roles of ASR error types (in 6.2) and LM predictions (in 6.3) in order to gain more insight concerning these complex questions and to refine the design of further follow-up experiments.

6.1. Human vs ASR system answers

The global human WER, computed on the central word of the transcribed stimuli, is 12% for the American English test and 15% for French. These rates take into account all stimuli, with the exception of the training and the distractor ones. When considering only the subset of correctly ASR-transcribed stimuli (0% system WER), a residual human error rate of about 1% is measured for both languages. On the complementary subset (100% system WER), the human WER increases to 16% for the English subjects and to 18% for the French ones. To measure the significance of this WER increase, the observed differences have been checked statistically. They are statistically significant for both American English (two-tailed t-test, $t(100)=10.293$, $p<.0001$) and French (two-tailed t-test, $t(76)=6.182$, $p<.0001$).

Humans are thus performing 5 to 6 times better than the ASR system on the speech chunks' central word set, for which the ASR system gave 100% misrecognized words. This human/machine WER ratio, on a reduced subset of difficult items, is lower than previous assessments[2],[3] which placed human transcription errors an order of magnitude lower than the best speech recognizer. Our measured ratio is closer to that of [4], and more recently to the one measured by W. Shen [8] who extended our methodology on various near-homophone English words.

6.2. WER vs type of error

The relationship between the ASR error taxonomy (cf. subsection 5.1) and the human performance was examined for both American English and French. Speech chunk transcriptions were separated according to the ASR error type on the target word. Correct items have also been considered. The statistical significance of the factor "type of the error" was checked for both languages. One-factor ANOVA analyses (using as levels the different types of errors as listed above and a *no-error* type for the correctly ASR-decoded stimuli) revealed that the measured factor is statistically significant for American English ($F(5,100)=18.6$, $p<.0001$) and French ($F(5,95)=23.9$, $p<.0001$) perceptual tests. The error type seems to play a role on the perceptual scores. Humans produce more errors for the stimuli for which the system missed not only the target word, but also the surrounding context (error type (5)), than for stimuli for which the only target word was deleted or inserted (error types (2) or (3)), the other surrounding words remaining correct by the ASR system. This finding, in line with the previous observations, further suggests that some stimuli are hard to transcribe both by ASR systems and humans. These stimuli contain intrinsic ambiguity.

6.3. Human error vs language model prediction

As argued in Section 2, none or only low information for the discrimination between (near)-homophones can come from the acoustics. Confusions are then left to be explained by the lan-

guage model, which produces inappropriate probability estimates for the involved word sequences.

To investigate the question of model bias *vs* intrinsic ambiguity, we focus on the subset of stimuli corresponding to the ASR near-homophone substitutions (error type (1)). Using respectively the English and French 4-gram language models, log-likelihoods (*llh*) have been computed for the transcriptions of the stimuli in both languages, including for each stimulus, the reference transcription and the one produced by replacing the central word by its near-homophone counterpart. These *llh* values are used by the ASR system when taking the word decision.

The hypothesis of contextual ambiguity may then be assessed via the *llh* values given with the language models. The *llh* difference between the reference transcription and the one with the near-homophone substitution gives a measurement of the stimulus' contextual ambiguity during ASR: large negative deltas are in favor of the reference transcription and thus are indicative of unambiguous stimuli. Ambiguity appears for close to zero deltas: here small differences arising from the acoustic modeling of near-homophones may play a role in the ASR decision. Finally clearly positive values go in favor of the near-homophone substitution. Table 3 shows some stimuli examples with the corresponding *llh* deltas. The question is then whether humans better transcribe unambiguous stimuli with respect to the *llh* delta or not. This ambiguity measure (*llh* deltas) was

Example stimuli	Δ *llh*	*Hum. ans*	*ASR ans*
ensuivie le vingt ***et*** *un avril*	-8.81	OK	OK
"caused the twenty-first of April"			
réaliste le plateau **est** *une coquille vide*	-1.19	err	OK
"realistic the stage is an empty shell"			
in six weeks ***and*** *Cisco systems the*	0	OK	OK
dans ce cas ***et*** *l'une des*	0.55	err	err
"in this case and one of the"			
fear of god **in** *many of them*	0.56	err	err

Table 3: Examples of stimuli in English and French with human and ASR system answers. Δ *llh* = *llh* difference between *REF* and *HYP* chunks. Ambiguity appears for close to 0 and positive Δ values.

correlated with the human answers. ANOVA statistical analyses with the single factor "LM prediction" ("ambiguous vs. unambiguous context", i.e. positive vs. negative deltas) were conducted for both languages. In French, human transcriptions correlate with the LM prediction, that is humans are better in correctly transcribing the speech chunks for which the reference word has been predicted. For American English this tendency was not observed. However for both languages the measured results are not statistically significant. Further experiments need to be designed, where the stimuli sets are selected using the delta *llh* as an additional control parameter.

7. Discussion

This contribution aimed at measuring human speech transcription accuracy in conditions simulating those of state-of-the-art ASR systems in a very focused situation. Sets of 7-word stimuli were extracted from American English and French broadcast news, containing a central word which potentially has an ASR system error. The major controling parameter for stimulus selection was the presence of a word which highly contributes to the ASR word error rate. The control parameter leads to the selection of highly frequent, monosyllabic, near-homophonic words: *and/in* in American English and *et/est* in French. The selected stimuli were presented to groups of native subjects. Average human transcription errors of 12% (repsectively 15%) were measured on the central word for English (respectively French). A contrastive statistical analysis on correctly vs wrongly ASR-decoded stimuli highlights that humans produce significantly more errors on stimuli misrecognized by the ASR system than on those correctly decoded by it. For the latter (0% ASR word error rate), a residual human word error rate of 1% is measured. These results are consistently observed for both languages with similar rates. Globally our study shows that the human error rate is 5 to 6 times lower than the ASR system in this limited context near-homophone transcription task. Informal analyses confirm that human transcription accuracy also varies with syntactic and semantic ambiguities, which were not the focus of this study.

Future experiments aim at clarifying the link between measured language model *llh* deltas on near-homophones and the corresponding perceived ambiguity in English and French. Finally, further investigations are planned to rank ideas aiming at reducing the model bias and the induced speech ambiguity. These include improved models with large context-dependent pronunciation options limiting near-homophony, and additional levels with some syntactic and semantic information.

8. Acknowledgements

This work was partially financed by OSEO under the Quaero program.

9. References

[1] Adda-Decker, M., "De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux", Journées d'Etude sur la Parole, France, 2006.

[2] Deshmukh, N.et al., "Benchmarking human performance for continous speech recognition", in Proc. of ICSLP, Vol.4, 1996.

[3] Lippmann, N., "Speech recognition by machines and humans", "Benchmarking human performance for continous speech recognition", Speech Communication, vol. 22, 99 1–15, 1997.

[4] Shinozaki, T. and S. Furui, "An assessement of automatic recognition techniques for spontaneous speech in comparison with human performance", in Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.

[5] Gauvain, J.L. et al., "Where are we in transcribing French broadcast news", in Proc. Interspeech, pp.1665–1668, 2005.

[6] Galliano, S et al., "ESTER PhaseII Evaluation Campaign for Rich Transcription and Broadcast News", in Proc. Interspeech 2005.

[7] Barras, C. et al., "Transcriber: development and use of a tool for assisting speech corpora production", Speech Communication, vol. 33(1-2), 2000.

[8] Shen, W. et al., "Two Protocols Comparing Human and Machine Phonetic Recognition Performance in Conversational Speech", Proc. Interspeech, 2008.

[9] Nemoto, R. et al., "Speech errors on frequently observed homophones in French: perceptual evaluation vs automatic classification", Proc. LREC, 2008.