Cross-lingual study of ASR errors: on the role of the context in human perception of near-homophones

I. Vasilescu, D. Yahia*, N. Snoeren*, M. Adda-Decker*,** and L. Lamel**

∗Spoken Language Processing Group, LIMSI-CNRS, 91403 Orsay
∗∗Laboratoire de Phonétique et Phonologie LPP-CNRS, UMR 7018, 75005 Paris
{ioana,yahia,nsnoeren,yahia,madda,lamel}@limsi.fr

Abstract

It is widely acknowledged that human listeners significantly outperform machines when it comes to transcribing speech. This paper presents a paradigm for perceptual experiments that aims to increase our understanding of human and automatic speech recognition errors. The role of the context length is investigated through perceptual recovery of small homophonic words or near-homophones yielding frequent automatic transcription errors. The same experimental protocol of varied size speech stimuli transcription is applied to both French and English. Our hypothesis is that ambiguity due to homophonic words reduces with context size for both languages, which in turn should entail reduced perception and transcription errors. The results show that context plays a central role as the human word error rate decreases significantly with increasing context. The long-term aim is to improve the modelling of such ambiguous items to reduce automatic errors.
Index Terms: automatic speech recognition errors, error analysis, linguistic variation, near-homophones, function words, lexical context, perceptual paradigm.

1. Introduction

During the last decade, several studies have established that humans significantly outperform machines on speech transcription tasks [3]. These observations are particularly true when large surrounding contexts (i.e. complete and long sentences) are available. These studies demonstrated that human listeners are better at handling many aspects of variation, such as pronunciation variants, noise, disfluencies, ungrammatical sentences, and accents, which remain challenging for current ASR systems.

An order of magnitude higher word error rates was reported for ASR systems as compared to 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 contextual information conditions of automatic speech decoders for human perception experiments. Stimuli comprising one target word embedded in a one word left/right context allowed to simulate word bigram networks as used by automatic decoders. In this very limited context condition, results indicated degraded human performances compared to previous studies [2, 3]. That is, instead of outperforming automatic recognizers by an order of magnitude, humans produce about half the errors of an automatic system. These studies highlight the importance of lexical context for accurate human transcription in that the information is not exclusively taken locally from the acoustic signal.

Such studies can benefit from the research in automatic speech recognition (ASR) which has fostered the development of very large scale speech corpora with corresponding orthographic transcriptions. Today's best ASR speech models are quite efficient, however they have not yet reached the status of being able to perfectly take into account all the sources of variability met in the spoken data (e.g. speech rate, stress, emotions, voice quality, health...). This study describes some of our ongoing research efforts on analyzing speech regions that are difficult for machines with the long-term aim of improving the ASR models.
Given these observations, we further perceptually explore the lexical context issue and its varying role in lexical decoding. A particular emphasis is put on *the context length in solving the local ambiguity*. The main question addressed is whether the context helps in disambiguating speech regions subject to ASR errors. In particular, do error rates increase with shorter contexts? A second question of interest concerns the influence of the *language*, i.e. may a similar effect be observed across languages?

The next two sections are dedicated to the working hypothesis (section 3) which motivates the experimental paradigm (section 2). In section 4 the perceptual design is introduced followed by the main results (section 5). The final paragraph is dedicated to discussion (section 6).

2. Perceptual Investigation of the ASR Errors

ASR transcription errors highlight speech regions which are problematic with respect to the ASR system's capacities. These speech regions correspond either to intrinsic ambiguities or to some type of variation not properly accounted for in the system's speech models. In any case, from an ASR perspective, transcription errors can be viewed as ambiguous speech regions with acoustical and/or contextual confusability. These ambiguities may either arise as a result of a simplified speech model (model bias), or be due to intrinsic spoken language ambiguities (language bias).

In [5] a perceptual paradigm was developed to identify a target word in 3-gram left and 3-gram right lexical contexts. Such 7-gram length stimuli (that is, 3 words left and right available to disambiguate a central target word) correspond to the maximum span of 4-gram language models typically used in ASR. The experiment was conducted in French and

American English on word pairs frequently misrecognized by the automatic systems. The Automatic Word Error Rates (WER) for the involved English word pairs *and/in* were found to be about 15% for all data, whereas they rose above 20% (*et/est)* in French broadcast news data [1].

The results showed that some lexical environments such 7-gram sequences do not provide sufficient information to disambiguate the central lexical targets. In particular, the results provided evidence that humans achieved significantly worse results on stimuli including ASR errors, than on stimuli which were correctly decoded by the automatic transcription system. Thus, a clear correlation in lexical transcription success (respectively failure) could be established between ASR systems and humans, in that the near-homophone ambiguity penalizes both the system and the human listener, even though humans seem to develop complementary strategies to overcome the local complexity. Results stressed the relevance of the *context* parameter as an operational tool.

3. Working Hypothesis

In this study, we make use of a new perceptual paradigm to explore the role of *increasing lexical context* in the disambiguation of near-homophone targets. The experiment involves similar stimuli to the previous experiment, namely lexical targets that are near-homophone short function words mostly prone to ASR errors [1]. The experiments are based on stimuli selected from large automatically transcribed speech corpora, with ASR results (presence or absence of ASR errors) as control parameters. In order to sort the language specific *vs.* independent effects, the experiment is conducted on French and English.
The target words are frequently misrecognized words, e.g. acoustically poor grammatical words likely to be confounded with corresponding near-homophones[1]. In the current experiment these words included *et, est, des, les, à, a* in French and *and, in, the, a, is, was* in English[2].
They are observed in various lexical environments corresponding to spoken regions that are erroneously transcribed by the automatic system, i.e., contexts where substitutions, deletions and insertions have been observed[3]. For each target word, embedding stimuli of length 3, 5, 7 and 9 words are extracted from the data, which allows simulation of a maximum language model span of 2-gram, 3-gram, 4-gram and 5-gram language models. Unigram stimuli are not retained to avoid potential word boundary segmentation problems on very short items. The perceptual paradigm requires human listeners to transcribe target words in contexts that vary from 3 to 9-grams, that is from a minimal context to one that is larger than those explored by most ASR systems and larger than the fixed 7-grams [5]. Recall that seven word contexts correspond to the maximum span of the ASR 4-gram language model.

[1]In [8] pseudo-homophony is defined as the inability to distinguish minimal pairs in L2 language which sound the same in L1 language of the speaker, e.g. wright/light. The definition is extended here to such lexical items which may "sound identically" for an ASR system as they differ in no more than two phonemes. Such acoustic proximity makes them near-homophones.

[2]The selected items caused the most frequent automatic errors, whether it is about deletions, insertions and mutual or with other words substitutions.

[3]http://www.itl.nist.gov/iad/mig/tests/rt/2002/software.htm

Type of stm. (+/- err.)	French	English
Substitutions	109 (54%)	124 (56%)
Insertions	29 (15%)	37 (17%)
Deletions	40 (20%)	41 (19%)
Correct	22 (11%)	18 (8%)

Table 1: Distribution of the excerpts across the tests according to the type of stimulus, i.e. erroneous and correct in French and English

4. Experimental Protocol

In the current experiment, the QUAERO French (2009) and English (2009 and 2010) test data are used (www.quaero.org). In both languages, data consist of various broadcast shows. In French, recordings feature mainly the standard version of the language. In English, shows come from British and American television channels.

4.1. Data and Stimuli selection

The selected audio excerpts correspond to n-grams which contain as a central item one of the target words mentioned above. A total of 200 excerpts containing central target words was selected from the French and 220 excerpts from the English test data. About 10% of the target items were correctly transcribed by the ASR system, the remaining target words either substituted, deleted or inserted.

Table 1 sums up the distribution of the correct and ASR erroneous excerpts within the test sets. The errors featured in the selected excerpts are equally balanced among the target words.

For each of the targets, 4 embedding stimuli of length 3, 5, 7, 9 words were extracted, resulting in a total of 800 audio stimuli in French and 880 in English. Table 2 illustrates the stimuli selection strategy. The selected stimuli feature the three types of errors in French and English and the four possible embedding contexts. The selection criteria result in four factors that are taken into account: *context size* (3, 5, 7, 9-grams), *automatic transcription of the target word* (i.e. correct vs. erroneous), *type of automatic error* (i.e. substitution, insertion, deletion) and *target word*.

4.2. Design and Participants

The stimuli were divided into 4 distinct sets of 200 (French) and 220 (English) stimuli, each set including all target words, but with stimuli of one of the context sizes selected randomly. Each stimuli set required a test population of at least 10 human transcribers: 40 participants completed the French test and 76 the English one, equally distributed across the four tests.

The rationale of this test design was to have each target word transcribed in its various embedding context length without repeating the same target word to the same human listener. The stimuli were presented for transcription through a web designed interface.

5. Perceptual Results

Human transcription performance was measured in terms of human WER and compared with the automatic solution for the central targets according to the factors mentionned here above. In total there are 8K transcriptions for French and more than 16K transcriptions for English, taken into account in the fol-

ASR err.typ.		French	English
SUB	REF	comme la région Auvergne EST légitime pour communiquer auprès la région Auvergne EST légitime pour communiquer région Auvergne EST légitime pour Auvergne EST légitime	so the review panel WAS headed by David Davis the review panel WAS headed by David review panel WAS headed by panel WAS headed
	HYP	comme la région Auvergne ET légitime pour communiquer auprès	so the review panel IS headed by David Davis
DEL	REF	pour cent alors que LES bénéfices explosent en plus cent alors que LES bénéfices explosent en alors que LES bénéfices explosent que LES bénéfices	pig remains will slip IN defeat for the first remains will slip IN defeat for the will slip IN defeat for slip IN defeat
	HYP	pour cent alors que * bénéfices explosent en plus	pig remains will slip * defeat for the first
INS	REF	investir dans le travail * investir dans l'entreprise dans le travail * investir dans l' le travail * investir dans travail * investir	the process of developing * real competitive market is process of developing * real competitive markets of developing * real competitive developing * real
	HYP	investir dans le travail A investir dans l'entreprise	the process of developing A real competitive markets is

Table 2: Examples of experimental design and selection of the stimuli. SUB=substitutions, DEL=deletions, INS=insertions; REF=manual transcription of reference; HYP=automatic solution

WER/n-gram	3-gram	5-gram	7-gram	9-gram
ASR correct (French)	7.3	1.8	2.3	0.9
ASR incorrect (French)	34	24	21	18
Global (French)	31	21	18	16
ASR correct (English)	5.8	2.5	1.9	1.7
ASR incorrect (English)	35	26	23	20
Global (English)	33	24	21	19

Table 3: Human WER for ASR erroneous and error free stimuli according to n-gram size.

lowing analysis by factor type.

5.1. The Impact of the Context Size

Table 3 displays the human performance in terms of WER on the central target word according to *stimulus length* (3, 5, 7, 9-gram). The factor *automatic transcription of the target word* (i.e. correct vs. erroneous) is equally considered. Human WER underline that target words elicit overall more difficulties: WER average 22% for each lexical item considered as the centre of the n-gram (21.5% for French and 22.5% for English). The effect is language independent. The result is consistent across the four sub-tests in both languages and allows to globally consider the transcriptions. It is also consistent with previous findings [5, 7].

Not surprisingly, human WER decrease with increasing context length for both perceptual experiments (figure 1). The results suggest that English stimuli posed more transcription difficulties than their French counterparts: WER for all n-grams are higher than the French ratios.

On the whole, human performance is correlated to context length for both languages, that is human listeners transcribe better when a larger context is available. For both languages, the benefit is particularly high when increasing the context from 3 to a 5-gram. Increasing the 7-gram context to a 9-gram one (recall that 7-grams correspond to the maximul span of the language model) does not result in a noticeable decrease in human errors. The results suggest that larger context than 9-grams is needed to come closer to error free solutions.

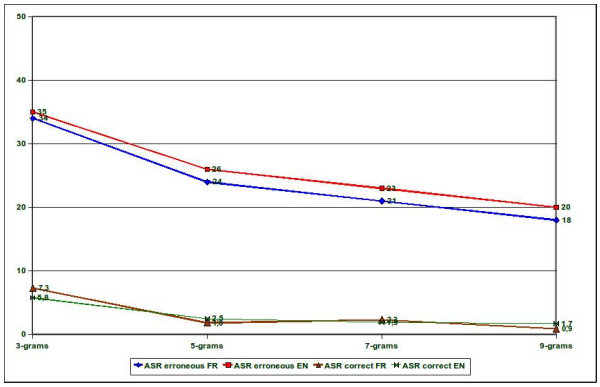

Figure 1: Human WER in French and English as function of n-gram size, for stimuli without and with ASR errors.

Finally, automatic and human errors show a positive correlation: the ASR erroneous regions lead to 24% human WER in French and 26% in English *vs.* 3% and 2% for the ASR error free stimuli. In all, the observed ratios suggests spoken regions erroneously transcribed by ASR system are also challenging for human transcribers.

5.2. The Impact of ASR Error Type

Human WER were computed on stimuli subsets according to ASR error typology, that is to substitutions (S), insertions (I) and deletions (D) of the target words for both languages. The ASR system developed at LIMSI [6] produced 23.8% WER for 2009 data and 23.7% and 17.31% for English 2009 and 2010 data respectively. The substitutions prevail (around 10 to 12%), followed by deletions (5 to 11%) and insertions (2%). Ratios are comparable in French and English. Table 4 illustrates the human error computed as a function of the ASR error typology. The two languages follow similar trends, that is human performance is correlated to the ASR error type. The perceptual error pattern that emerges in both languages corresponds to the observation that words subject to automatic deletions yielded the highest human WER. This result support the hypothesis that

Human WER	S	I	D
French	24	14	32
English	26	13	37

Table 4: Human WER according to the type of automatic error, i.e. substitution (S), insertion (I), deletion (D).

French target words	a	à	et	est	des	les
% WER (%)	24	24	23	14	22	23
English target words	in	and	a	the	is	was
% WER (%)	30	26	18	20	26	31

Table 5: Human WER displayed as a function of central target words in French and English

poor acoustic information may be a challenge for both ASR system and humans. In opposition to substitutions or deletions, automatically inserted words yielded the lowest human WER which suggest that they occur principally in less ambiguous contexts. As a general observation, human errors vary with type of automatic substitution. For instance, contexts which lead to ASR mutual substitutions between true homophones such as *est*(to be, is)/*et*(and) in French are also more problematic for human transcribers than stimuli containing other ASR type of errors.

5.3. The Impact of the Central Target Word

The present experiments focused on short ambiguous function words, homophones or near-homophones which are common errors in ASR transcriptions. These words included *et, est, des, les, à, a* in French and *and, in, the, a, is, was* in English. WER computed according to the target word are shown in table 5 giving human performance as a function of the central item. Error patterns in French show that WER are equally distributed among target words. The word *est* (to be, is) is an exception as the WER are below the ratios observed for the other items. A closer look to the data shows that most of the human erroneous transcriptions involve the confusion *c'est* (this is) *vs. ce* (this) and the mutual confusion *est* (to be, is) *vs. et* (and), whereas the remaining contexts are less ambiguous. In English WER for the target words range from 18% (*a*) to 31% (*was*). The results suggest that word size does not influence WER ratios *per se*, that is short items are not necessarily more ambiguous than longer ones (e.g., *a vs. was*).

6. Discussion

In the present contribution we applied a recently proposed paradigm for perceptual experiments to investigate human decoding capacities on ASR error speech stimuli. The paradigm was designed to assess human speech transcription accuracy in conditions simulating those of state-of-the-art ASR systems in a very focused situation. We investigated the most commonly observed errors in automatic transcription, namely the confusion between, and more generally speaking the erroneous transcription of near-homophonic words in French and English, and evaluated these in a series of perceptual tests involving human transcribers.

The results confirm that human listeners performed on average 5 to 6 times better than the ASR system on the speech chunks' central word set. The perceptual tests also show that speech errors are typically modulated by a number of factors. The context plays a central role as the human WER decrease significantly with increasing context. In particular, extending the context from 3 to 5-gram brings the most significant improvement (10%). However, 9-grams, i.e. one-word larger contexts than the maximum span of the language model (7-grams), do not provide sufficient contextual information to outperform human transcription performance (18% WER). The observed effects are language-independent. Human listeners produced also significantly more errors on stimuli misrecognized than on those correctly decoded by the ASR system, where a residual error rate of about 4% was measured. Ongoing work concerns statistical in-depth analyses of the perceptual results.

Future investigations are planned to reduce the model bias so as to better cope with speech ambiguity. These include models with large context-dependent pronunciations limiting near-homophony, as well as syntactic and semantic information.

7. Acknowledgements

This work is partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation.

8. References

[1] Adda-Decker, M., "De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux", in Proc. of JEP, 2006.

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

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

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

[6] Lamel, L., Quaero Program - CTC Project - Progress REport on Task 5.1: Speech to Text, CD.CTC.5.6., Quaero Program, 2010.

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

[5] Vasilescu. I. et al., "A perceptual investigation of speech transcription errors involving frequent near-homophones in French and American English", in Proc. of Interspeech, Brighton, UK, 2009.

[8] Cutler, A., 2005, The lexical statistics of word recognition problems caused by L2 phonetic confusion, Proc. of Interspeech, Lisbon, Portugal, 2005.