Cross-lingual studies of ASR errors: paradigms for perceptual evaluations

I. Vasilescu[1], M. Adda-Decker[1,2], L. Lamel[1]

[1]LIMSI-CNRS, [2]LPP-CNRS,
[1]B.P. 133 91403 Orsay France, [2]UMR 7018, 19 rue des Bernardins, 75005 Paris
ioana@limsi.fr, madda@{univ-paris3,limsi}.fr, lamel@limsi.fr

Abstract

It is well-known that human listeners significantly outperform machines when it comes to transcribing speech. This paper presents a progress report of the joint research in the automatic vs human speech transcription and of the perceptual experiments developed at LIMSI that aims to increase our understanding of automatic speech recognition errors. Two paradigms are described here in which human listeners are asked to transcribe speech segments containing words that are frequently misrecognized by the system. In particular, we sought to gain information about the impact of increased context to help humans disambiguate problematic lexical items, typically homophone or near-homophone words. The long-term aim of this research is to improve the modeling of ambiguous contexts so as to reduce automatic transcription errors.

Keywords: ASR, HSR, speech transcription errors.

1. Introduction

Nowadays an increasing number of innovative applications make useful automatic speech transcription in particular to access multimedia material. Automatic speech transcription is thus applied to improve meaning-based access to multimedia collections containing spoken content: subtitling of videos, search for precise portions of audio-visual archives, automated reports of meetings, extracting and structuring of information (Speech Analytics) in multimedia contents (Web, call centers, . . .).

However, transcription errors persist, which are more or less problematic depending on the application. For instance, information retrieval is relatively tolerant to errors (up to 30%), but systematic errors on certain named entities can be prohibitive. On the contrary, subtitling or meeting transcription have a very low tolerance to errors, and even low word error rates (below 5%) are too high for the end-users. Consequently, research on error diagnosis and classification are still important to improve the current stat-of-the-art large and very large vocabulary continuous speech recognition systems.

One of the strategies for error analysis consists to carry out perceptual tests with human listeners. It is widely acknowledged that human listeners significantly outperform machines when it comes to transcribing speech, as current *automatic speech recognition (ASR) systems* still have difficulty handling all sources of variability conveyed by the speech signal. Therefore, bridging the gap between humans and machines by taking advantage of perceptual strategies has become an active research area.

This paper first provides an evaluation report of the studies in human/machine comparison on speech transcription tasks over the last decade and then sums up our own experiments on frequent ASR misreconized words explored through various experimental paradigms in French and English mainly developed under the Quaero Programme (`www.quaero.org`).

2. Automatic vs Human speech transcription: a state-of-the-art

Today's automatic speech recognition (ASR) systems have difficulty to adjust efficiently to the sources of variability arising from the vocal profile of the speakers and/or contextual patterns of the verbal interaction (including among others, coarticulation effects, pronunciation, disfluencies, speaking style, gender or age related patterns, physiology, emotions, foreign and regional accents etc.). Elements of this kind may highly impact the transcription task (Benzegiba et al., 2007). Recent improvements seen in state-of-the-art ASR systems are mostly due to enhanced discriminative training schemes combined with increased amounts of training data and computing power. What's more, the present ASR transcription accuracy levels still fall short of human performance: extending training data may not be the solution to come up to the human level of accuracy (Moore, 2003).

Studies on *human speech recognition (HSR)* highlight that human transcription of spoken data remains unquestionably superior to most performant automatic results, even though the gap between the humans and the machines may reduce significantly on some focused conditions (Deshmuk et al., 1996a). HSR is generally flexible, robust and efficient in the face of a large variety of distortions, both experimentally applied and naturally occurring, whereas under similar conditions ASR systems fail. Several attempts have been conducted to compare humans and machines on various speech recognition tasks (Lippmann, 1997; Moore and Cutler, 2001; Pols, 1997; Scharenborg, 2007). These studies show that humans remain up to one order of magnitude better than machines in recognizing speech as evaluated in (Lippmann, 1997) and confirmed one decade later in (Scharenborg, 2007), both authors providing results from various comparative experiments and presenting human and automatic word error rates (WER) for a wide range of tasks and conditions.

The claim that automatic speech recognition performance lags about an order-of-magnitude behind human performance still remains true.

2.1. Automatic vs Human speech transcription: levels of comparison

HSR performances have been compared to the automatic ones at various speech levels, from articulatory features and phonemes to sentences. Different types of audio material have been covered, from clean to degraded speech, and with respect to various amounts of training data. The purpose of such comparative studies is to underline what it is that makes human speech recognition so much superior to machine speech recognition and what can be learned from human speech recognition to improve the automatic performance. Conclusions drawn from various perceptual paradigms lead to the following assessment: human listeners seem to have multiple sources of information available that are unavailable for ASR systems (*e.g.* ”higher-level” knowledge which is not incorporated in statistical models currently used in the field of the automatic recognition).

Human performance on speech transcription tasks is particularly high when the comparison with the automatic output is drawn on excerpts benefiting from large surrounding contexts (i.e. complete and long sentences) (Lippmann, 1997). However, even when ”higher-level” information is removed, human listeners are still better at transcribing stimuli suggesting the use of different and more efficient cues during recognition (Shinozaki and Furui, 2003). The ”lower-level” information employed by humans may also be different: most likely, listeners make use of all the information present in the acoustic signal whereas ASR systems take advantage only of the information encoded in the acoustic features (Scharenborg, 2007).

2.2. Automatic vs Human speech transcription: cross-fertilization approaches

There is a recent growing interest in a cross-fertilization between the domains of human and automatic speech recognition. The two research fields are effectively closely related, yet they are led by different aims (Scharenborg, 2007). Human and automatic speech recognition are both concerned with understanding the process of extracting linguistic information from the acoustic signal (Moore and Cutler, 2001). In the HSR framework, the goal is to understand how we as listeners recognize spoken utterances, whereas ASR is concerned with building algorithms that are able to recognize the words in a speech utterance automatically, under a variety of conditions, with the least possible number of recognition errors.

Improving the ASR performances thanks to techniques highlighted by human strategies in successfully transcribing speech is thus a recurrent topic in both HSR and ASR studies. With respect to the improvements of ASR techniques thanks to HSR achievements, two research paradigms may be mentioned :
(i) a first approach is *to take advantage of the theoretical models developed in HSR to improve the ASR methods*;
(ii) a second approach consists in *questioning the human performance in conditions which stick as close as possible to current ASR architecture* to gather the missing information from the latter which may be ”injected” in the ASR systems.

The *first approach* is illustrated by research as detailed in (Scharenborg et al., 2005), who describes a model of human speech recognition built using techniques from ASR. Adopting a unique evaluation metric to compare the performance of humans and automatic systems in recognizing the same spoken data as in (Cutler and Robinson, 1992) is another possible way. The authors adapt the response time as the HSR evaluation technique to evaluate both human and automatic performances.

The second approach is concerned with focused comparisons between HSR and ASR, *via* perceptual experimentation with the aim of pointing out quantifiable recognition strategies to further ”inject” into ASR technology. However, integrating knowledge from human speech recognition into ASR systems is a hard task and such comparisons often led to the frustrating conclusion that it is unclear which human winning strategy for recognizing speech would be relevant for improving ASR systems. What is unquestionable, is that humans are more flexible than ASR systems (Pols, 1997), and able to deal successfully with spoken data varying in speech style and type or listening environment.

It is worth noting that most of these studies have focused on end-to-end comparisons of the recognition process, i.e. how do humans and machines perform on complete spoken utterances. As a typical example, an order of magnitude higher word error rates was reported for automatic speech recognition 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 (Deshmuk et al., 1996b). A similar difference in performance between humans and automatic decoders has been reported for various sources featuring prepared or spontaneous speech (Leeuwen et al., 1995; Lippmann, 1997).

There are less studies on more controlled discrimination tasks making use of spoken stimuli with length inferior to the utterance and/or syntagm level (Scharenborg, 2007). The aim of the studies focusing on recognition tasks of excerpts below the sentence level is to better decompose the human performances and identify efficient recognition strategies through more controlled stimuli subsets. Such studies presuppose that in order to conduct a fair human-machine comparison both sides should utilize the same/similar types and/or levels of information. Several studies show that the WER effectively converges in testing setups where humans could not use supplementary cues, such as the context of the conversation, the grammar of the spoken language or certain words (Lippmann, 1997). However, human perception still remains performant during experiments which stick closer to the conditions of ASR systems processing by reducing the amount of information available to transcribe a target item in terms of context or level of usable information (from acoustic to higher-level that is syntactic or semantic). Human listening appears to be performant at various unit levels, e.g. words (Lippmann, 1997), logatomes (Meyer and Wesker, 2006), consonants and vowels (Cutler and Robinson, 1992; Sroka and Braida, 2005). In such restricted conditions compared to sentence level experiments, human perception still remains equal to or outperforms the automatic output. However, the gap between humans and machines decreases showing that missing information affects both. Human listeners do not always recognize everything correctly, and even when they do, they find some items more difficult to process than others. To illustrate this point, in (Meyer and Wesker, 2006) the human vs machine comparison was conducted on logatoms in order to ensure equal recognition conditions: automatic errors remain 30% above that of humans even though no contextual knowledge can be exploited. Human listeners seem to use some phonetic features such as voicing to overcome the lack of linguistic information. Another study conducted by (Shinozaki and Furui, 2003) 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 simulation of the word bigram networks used by automatic decoders. In this very limited context condition, results indicated degraded human performance: they produce about half the errors of an automatic system. Finally, in (Shen et al., 2008) native speakers of Italian (L1) were asked to orthographically transcribe Japanese and Spanish speech (L2): the choice of languages was designed to maximize phonetic overlap and minimize cross-language phonotactic mismatches (defined as systematic cross-language biases that lead to errors when subjects of one language attempt to transcribe data from another). Humans provided 15% more accurate results than an ASR-based phone recognizer.

2.3. Automatic vs Human speech transcription: conclusions and lessons

Studies on human vs automatic transcription of speech have pointed out the sharpness of the human perception in decoding spoken data. Such comparisons moved over time from spoken data significantly below the syntagm level, that is consonants and vowels, to entire and long sentences. Various ”winning” strategies have been as well observed which help humans performing up to one order of magnitude better than machines. However, formalizing and then ”injecting” human strategies into ASR systems remains an objective complex to achieve. One may find a reason in the complexity of human decoding strategy of a given ambiguous item, involving simultaneously ”high” and ”low-level” information. Human also uses its own ”knowledge of the world”, which may go from the topic of the conversation to wider models which characterize a domain or life situation. The perceptual paradigms developed in our laboratory have a slightly different yet close aim: to benefit from the research in automatic speech recognition for (in)validating linguistic hypotheses that could provide answers to current questions in the ASR field.

3. Paradigms for perceptual studies

The perceptual investigations conducted at LIMSI rely on the following assumptions: ASR transcription errors highlight speech regions which are problematic with respect to the ASR system’s decoding capacities. *ASR transcription errors can be viewed as ambiguous speech regions with acoustical and/or contextual confusability.* Various reasons may be effectively identified which explain the occurrence of automatic speech transcription errors such as ASR system configuration (e.g. acoustic models, pronunciations, etc.), speaking style (read, prepared or spontaneous, speech quality (e.g. SNR), speaker production (accents, fluency, interactivity, register, etc.).

A comparison of human transcriptions with those of ASR systems, may then be indicative of either **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 (**language bias**) of the speech signal or to the **model bias**. ASR systems then offer opportunities to imagine innovative tools for the design of perceptual experiments to sort out the respective roles of model and language biases.
Perceptual experimentation described in this paper also aims at estimating the optimal additional information required both by human perception and by 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.

We focus here on the (near-)homophony[1] as source of frequent speech transcription errors.
(Near-)Homophony concerns not only words which are pronounced the same (e.g. fair, fare), but also words and words sequences words and words sequences and may arise from poor articulation, misplaced word boundaries, alternate pronunciations for words, noisy environments etc. Short (acoustically poor) and frequent (mostly function words) lexical items are the most often prone to confusions based on similar acoustic characteristics: they are acoustically poor, misarticulated as often repeated and occurring in shared contexts thus likely to be substituted by a large class of similar items. Consequently they are at the top of the word error list.

[1] In (Cutler, 2005) 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.

4. Experiments on near-homophone targets in fixed length contexts

In a first experimental work we investigated the perceptual discrimination of frequently misrecognized words such as short grammatical items with (near)-homophone pronunciations (Vasilescu et al., 2009).

4.1. Experimental design

For this experiment the following data have been employed: (i) for English the study made use of a subset of the NIST HUB4 corpus consisting of broadcast news shows from different radio stations (VOA, ABC, etc.), in total 2.5 hours corresponding to 24.7k words; (ii) for French a subset of the TECHNOLANGUE-ESTER corpus has been employed, in total 10 hours of broadcast news data corresponding to 94k words. The WER rate averaged 11.5% for both languages.

Perceptual experiments in French and English aimed at identifying a target word in 3-gram left and 3-gram right lexical contexts. Target words are near-homophone pairs frequently misrecognized by the ASR system, that is *est/et* for French and *and/in* for English. Table 1 shows examples of spoken excerpts selected for the perceptual evaluation. Such 7-gram length stimuli (that is, 3 words left and right available to disambiguate the central target word) correspond to the maximum span of 4-gram language models typically used in ASR. Such 7-gram stimuli were presented to populations of native listeners of both French and English and the human transcription of the central lexical item was compared to the reference and the automatic transcriptions. In total 129 chunks for English and 83 for French have been proposed to groups of 40 listeners in each language. Chunks contained target words erroneously transcribed by the automatic system (i.e. on which the system produced either a substitution, deletion or insertion) and and also excerpts correctly transcribed by the ASR system for which the automatic hypothesis corresponds to the human reference. Among the stimuli, 90% contain an ASR error (the remaining chunks have been correctly transcribed by the automatic system).

4.2. Results

The human performance have been measured for the central target word. Results showed that for some lexical environments such 7-gram sequences do not provide sufficient information to disambiguate the central lexical targets. The global human WER, computed on the central word of the transcribed stimuli, results in 12% for the American English test and 15% for French. Humans are thus performing 5 to 6 times better than the ASR system on the speech chunks' central word set. 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 (for the latter a residual error rate of 1% has been observed). A clear correlation in lexical transcription success (respectively failure) could be established between ASR systems and humans.

English	
1	REF: of the day **and** it is almost HYP: of the day **in** it is almost
3	REF: escape on tape ******* the two were in HYP: escape on tape **and** the two were in
French	
2	REF: politique aujourd'hui **est** essentiel d'approfondir HYP: politique aujourd'hui ******* essentiel d'approffondir
4	REF: de mai difficile **et** les syndicats HYP: de mai difficile **mais** les syndicats

Table 1: Examples of 7-words speech excerpts with transcription errors. ASR insertions/deletions are marked by ***. (Near)-homophone target words are **and, in** in English and **et, est** in French.

These results also suggest language-independent patterns as similar trends are noticed for English and French. The human transcriptions also varied with syntactic and semantic ambiguities.

Finally, these preliminary results stress the relevance of the *context* parameter the information being not exclusively grasped locally from the acoustic signal. In the particular case of (near-)homophone items the context is of a great importance as it carries cues which help to disambiguate such words. As a consequence, further experiments were designed to sort out the contribution of the lexical *context* in the disambiguation of local homophony.

5. Experiments on near-homophone targets in variable context size

The second phase of the perceptual investigation considered the role of *increasing lexical context* in the disambiguation of near-homophone targets (Vasilescu et al., 2011) . The experiments were based on the hypothesis that ambiguity due to homophonic words reduces with context size, which in turn should entail reduced perception and transcription errors.

In these experiments, the QUAERO French (2009) and English (2009 and 2010) test data are used. In both languages, data consist of various broadcast shows, not only just news. In French, recordings feature mainly the standard version of the language. In English, shows come from British and American television channels.

The target words are frequently misrecognized words, e.g. acoustically poor grammatical words likely to be mistaken with corresponding near-homophones. The list of near-homophone pairs was extended here to *et, est, des, les, , a* in French and *and, in, the, a, is, was* in English. They were 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.

5.1. Experimental design

5.1.1. Data selection and experimental setup

A total of 200 excerpts containing central target words was selected from the French and 220 excerpts from the English test data. The number of excerpts per language depends on the frequency of the three types of automatic errors as currently estimated in the ASR performances evaluation within the selected data. The WER as primary evaluation metric largely employed in automatic speech recognition is here taken into account and stimuli are selected according to the amount of substitutions, deletion and insertions concerning the most frequently misrecognized words. In the end, about 10% of the selected target items were correctly transcribed by the ASR system, the remaining target words either substituted, deleted or inserted. For each of the targets, 4 embedding stimuli of length 3, 5, 7, 9 words were extracted, and distributed at random into four different perceptual test sets. Recall that seven word contexts correspond to the maximum span of the ASR 4-gram language model (Shen et al., 2008; Vasilescu et al., 2009)

WER for the involved English word pairs were about 15% and above 20% in French broadcast data (Adda-Decker, 2006). 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.

Each test set was finally composed of stimuli of various context lengths, including each of the 800/880 (French/English) stimuli once and only once, in either a 3, 5, 7 or 9 word segment.

5.1.2. Test population

Each stimuli set required a test population of at least 10 human transcribers: 40 participants completed the French test and 76 the English one[2]. 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.

5.2. Human transcription processing

Erroneous stimuli according to ASR solution have been selected to portray one of the three types of automatic errors: substitutions, deletions or insertions. As for the human errors, they partly follow the automatic transcription trends. To take into account the human transcription proper, the following case figures are counted as errors: (i) the absence of transcription (global deletion); (ii) the partial deletion of the n-gram to transcribe including the target word; (iii) the erroneous transcription of the target word even though the error is different from the automatic one (e.g. the system deleted the target word but the human substituted the same target word).

WER/n-gram	3-gram	5-gram	7-gram	9-gram
ASR correct (French)	0.7	1	2.5	0.5
ASR incorrect (French)	37.3	24.9	19.7	17.5
Global (French)	34.2	22.5	17.9	15.7
ASR correct (English)	4.5	2.5	0.3	0.9
ASR incorrect (English)	40.4	26.6	23.2	19.9
Global (English)	37.6	25	21.1	18.2

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

6. Results

In the next sections, the human transcription performance as measured in terms of human WER for the central targets is discussed and compared with the automatic solution. The results are considered according to the following factors which motivated the perceptual paradigm configuration: *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*.

6.1. The factor context size

Firstly, human WERs underline that target words elicit overall transcription difficulties for both languages: global WER average is 24% for the target words (22.6% for French and 25.6% for English). The result is consistent with previous corresponding experimentations (Vasilescu et al., 2009; Shen et al., 2008) and corroborate the trends highlighted by the present state-of-the-art: the human perception leads to more efficient solutions in the case of some ambiguous spoken excerpts however both the human listeners and the ASR systems fail on some contexts particularly problematic. 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 also considered[3].

Results show that for each context size the correct transcription rate is above chance for both languages (χ^2 test, df=3, p$<$0.001). A one-factor ANOVA statistical analysis has been conducted for the human ratings to check the effect of the *stimulus length* factor. The measured factor is statistically significant for both French (F(3,8006)=80.391, p$<$0.001) and English (F(3,16412)=163.19, p$<$0.001). Figure 1 illustrates human WER according to n-gram size (3, 5, 7, 9-grams), ASR erroneous vs. correct stimuli and language. It is worth noting that the WER decrease with increasing contexts in particular for the ASR erroneous stimuli. The result support the hypothesis of the role of the context in the local disambiguation of (near-)homophone targets.

[2] A transcriber results are selected if the entire test has been completed. In English some test sets ”attracted” more participants than others and the overall participation exceeded 40 transcribers: for the current analysis all 76 answers are kept as being complete.

[3] Statistical analysis has been conducted with the R package `http://CRAN.R-project.org/`

ASR err.typ.		French	English
SUB	REF	comme la région Auvergne EST légitime pour communiquer auprès	so the review panel WAS headed by David Davis
		la région Auvergne EST légitime pour communiquer	the review panel WAS headed by David
		région Auvergne EST légitime pour	review panel WAS headed by
		Auvergne EST légitime	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	pig remains will slip IN defeat for the first
		cent alors que LES bénéfices explosent en	remains will slip IN defeat for the
		alors que LES bénéfices explosent	will slip IN defeat for
		que LES bénéfices	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	the process of developing * real competitive market is
		dans le travail * investir dans l'	process of developing * real competitive markets
		le travail * investir dans	of developing * real competitive
		travail * investir	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 stimuli selection. SUB=substitutions, DEL=deletions, INS=insertions; REF=manual transcription of reference; HYP=automatic solution

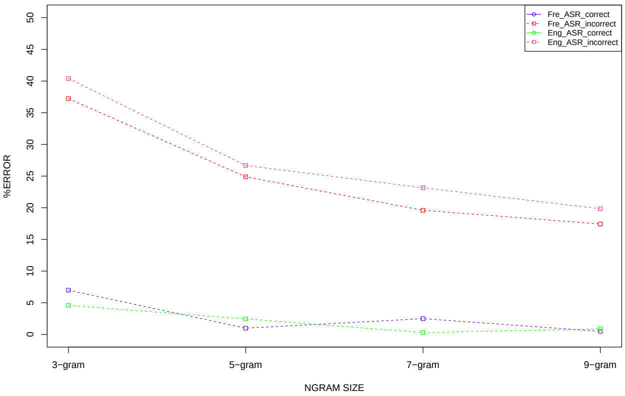

Figure 1: Human WER according to **n-gram size** (3, 5, 7, 9-grams), **ASR erroneous vs. correct** stimuli and **language**.

Human WER (%)	Sub	Ins	Del
French human WER	64	10	25
English human WER	61	12	26

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

6.2. The factor type of automatic error

Human WER were also computed on stimuli subsets according to ASR error typology, that is to substitutions (S), insertions (I) and deletions (D) of the target words by the system for both languages. The ASR system developed at LIMSI (Lamel, 2010) 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. The stimuli selection follows the ASR trends: test sets include mainly substitutions (64% French and 61% English), then deletions (25% and 26%) and finally insertions (10% and 12%).

Table 4 illustrates the human error computed as a function of the ASR error typology and takes into account all the context sizes. The two languages follow similar trends[4]. In opposition to substitutions or deletions which yielded relatively high human WER, especially for S errors, automatically inserted words yielded the lowest human WER which suggest that they occur principally in less ambiguous contexts.
The factor *type of automatic error* has been checked with one-factor ANOVA analysis. The factors is statistically significant for both languages, that is French (F(3,8006)=94.89, p<0.001) and English (F(5,16410)=44.105, p<0.001).

6.3. The factor target word

The present experiment 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. Target words are balanced across test sets (16% per target word in average for both languages).

WER computed according to the target word are shown in table 5 giving human performance as a function of the central item (the % take into account all the context sizes). Error patterns in French show that WER are globally equally

[4] It is important to mention that the Table 4 provides a general overview of the human difficulties in processing some sequences subject to substitutions, insertions or deletions without specifying the type of errors the humans produced, e.g. automatic substitutions did not lead systematically to human substitutions when a perceptual error occurred.

French target words	a	à	et	est	des	les
% WER (%)	19.3	19.9	18.2	11.8	15.6	14.9
English target words	in	and	a	the	is	was
% WER (%)	19.3	19.9	18.2	11.8	15.6	14.9

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

distributed among target words, the word *est* being an exception, which may be linked with some language specific sintactic/semantic constraints (e.g. the frequent occurence of the sequence *c'est* non-ambiguous for humans). In English, 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*). The factor *target word* appears to be statistically significant (one-factor ANOVA for French (F(3,8006)=10.61, p$<$0.001) and English (F(5,16410)=44.105, p$<$0.001) data).

7. Discussion

This paper presented a progress report of the joint research in the automatic vs human speech transcription as illustrated by experimental research conducted over the last decades. Comparisons between humans and automatic systems on various speech transcription tasks have shown that the former performed up to one order of magnitude better than machines. However, ”injecting” human strategies in ASR systems still remains a most challenging objective. In the present contribution we also 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 perceptual tests have been showing that speech errors are typically modulated by a number of factors, the context being a significant parameter for both the human and the ASR system. More extensive studies including larger data and based on a typology of sources of the local ambiguity are planned to reduce the model bias, and the induced speech ambiguity. These include models with large context-dependent pronunciations limiting near-homophony, as well as syntactic and semantic information.

8. Acknowledgments

This work is partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. We would like to thank Bianca Vieru for her precious help in statistical analysis, Dahbia Yahia and Natalie Snoeren for their help in building the perceptual test and interpreting the results. We are also very grateful to the 76 students of University of Philadelphia for their help in the English transcription test.

9. References

M. Adda-Decker. 2006. De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Proc. Journés d'Etude sur la Parole JEP 2006*, Dinard, France.

M. Benzegiba, R. De Mori, O. Deroo, T. Erbes, D. Jouvet, L. Fissore, P. Laface P., A. Mertins, C. Ris, R. Rose, V. Tyagi, and C.J. Wellekens. 2007. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(2007):763–786.

A. Cutler and T. Robinson. 1992. Response time as metric for comparison of speech recognition by humans and machines. In *Proc. of ICSLP*, pages 189–192, Banff, Canada.

A. Cutler. 2005. The lexical statistics of word recognition problems caused by l2 phonetic confusion. In *Proc. of Interspeech*, Lisbon, Portugal.

N. Deshmuk, R.J. Duncan, A. Ganapathiraju, and J. Picone. 1996a. Benchmarking human performance for continuous speech recognition. In *Proceedings of ICSLP 1996*.

N. Deshmuk, R.J. Duncan, A. Ganapathraju, and J. Picone. 1996b. Response time as metric for comparison of speech recognition by humans and machines. In *Proc. of ICSLP*, Philadelphia, USA.

L. Lamel. 2010. Quaero program - ctc project - progress report on task 5.1: Speech to text. Cd.ctc.5.6., Quaero Program, Limsi Orsay.

D.A. Van Leeuwen, L.-G. Van den Berg, and H.J.M. Steeneken. 1995. Human benchmarks for speaker independent large vocabulary recognition performance. In *Proc. of Eurospeech*, pages 1461–1464, Madrid, Spain.

N. Lippmann. 1997. Speech recognition by machines and humans. *Speech Communication*, 22(1997):1–15.

B. Meyer and T. Wesker. 2006. A human-machine comparison in speech recognition based on a logatome corpus. In *Proc. of Workshop on Speech Recognition and Intrinsic Variation*, Toulouse, France.

R.K. Moore and A. Cutler. 2001. Constraints on theories of humans vs. machine recognition of speech. In *SPRAAC Workshop on HSR as Pattern Classification*.

R.K. Moore. 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of Eurospeech*, pages 2582–2584, Genève, Suisse.

L.C.W. Pols. 1997. Flexible human speech recognition. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*.

O. Scharenborg, D. Norris, L. ten Bosch, and J.M. McQueen. 2005. How should a speech recognizer work? *Cognitive Science*, 29(2005):867–918.

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

W. Shen, J. Olive, and D. Jones. 2008. Two protocols comparing human and machine phonetic discrimination performance in conversational speech. In *Proc. of Interspeech*, Brisbane, Australia.

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

J. Sroka and L.D. Braida. 2005. Human and machine consonant recognition. *Speech Communication*, 45(2005):401–423.

I. Vasilescu, M. Adda-Decker, P. Hallé, and L. Lamel. 2009. A perceptual investigation of speech transcription errors involving frequent near-homophones in french and american english. In *Proc. of Interspeech*, pages 144–147, Brighton, UK.

I. Vasilescu, D. Yahia, N. Snoeren, L. Lamel, and M. Adda-Decker. 2011. Cross-lingual study of asr errors: on the role of the context in human perception of near-homophones. In *Proc. of Interspeech*, Firenze, Italy.