TOWARDS EXPLORING LINGUISTIC VARIATION IN ASR ERRORS: PARADIGM AND TOOL FOR PERCEPTUAL EXPERIMENTS

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

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

ABSTRACT

It is well-known 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 automatic speech recognition errors. The paradigm asks human listeners 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. The long-term aim of the this research is to improve the modeling of ambiguous items so as to reduce automatic transcription errors. To this extent we have been developing a tool, the Q-ERROR graphical interface, to facilitate the analysis of automatic speech recognition errors. As previous research has shown, speech recognition errors are often modulated by a number of factors, and it can be difficult to assess the impact of each. By enabling a user to filter data in large corpora, the proposed interface can also be used to help select the relevant stimuli for human perceptual tests.

Index Terms: automatic speech recognition, ASR errors, linguistic variation, acoustic-phonetic studies, perceptual test, perceptual paradigm, error analysis.

1. INTRODUCTION

Automatic speech recognition has fostered the development of very large scale speech corpora with corresponding orthographic transcriptions. The acoustic model training process generates segmentations into words and on a subword level, into phone segments. Depending on the pronunciation dictionary's options and the acoustic model's accuracy, the resulting phone stream provides a more or less accurate phonetic or phonemic labeling. Beyond enabling and optimizing voice-driven technologies, these annotated corpora are highly valuable resources for a wealth of phonetic or more generally linguistic studies [1].

Our long-term goals aim to increase both our knowledge of speech variation and to specify potential shortcomings in speech models used in state-of-the-art of Automatic Speech Recognition (ASR) systems. ASR systems make use of a speech model (typically composed of acoustic, pronunciation and lexical n-gram models) to decode an incoming speech stream. If there were no errors, the speech model could be considered as perfect, or at least as sufficiently informed to resolve all upcoming ambiguities. However, decoding errors do occur. The conventional speech model is still not to blame if errors arise due to intrinsic language ambiguities, which can be correctly solved only with additional semantic and pragmatic knowledge. In general, however, errors are also related to the speech model. There may be pieces of information missing in the model (e.g. acoustic models limited to neighboring phonetic contexts, the absence of some contextual factors, such as speech rate, stress, emotion, voice quality, health...). Language n-gram models may also poorly generalize to unseen contexts, whereas pronunciation dictionaries may miss some variants.

Although todays best ASR speech models are quite efficient, they have not yet reached the status of being able to perfectly take into account all observed acoustic variation. In the following, we will describe our ongoing research efforts on analyzing speech regions including ASR errors. This material is explored both perceptually and by specific acoustic and prosodic analyses. Results then need to be compared to error-free reference material. We expect this line of research to shed new light on less described acoustic variation factors in speech as well as on underlying mechanisms that allow humans and/or future ASR systems to cope with the variation.

In the following section an overview of some studies comparing ASR and human transcription performance is given. We then propose a new paradigm for perceptual experiments based on automatic ASR transcription in Section 3 and describe a preliminary perceptual experiment involving ASR errors. Section 4 introduces the new graphical interface being developed for ASR error analysis and for perceptual tests' stimulus selection. The major steps of the described work are summarized in Section 5.

2. COMPARING HUMAN AND ASR PERFORMANCE

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 (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, while these still remain important challenges 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 (in comparison to previous studies [2, 3]): 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. Given these observations, we would like to further look into the lexical context issue and its varying role in lexical decoding.

3. PARADIGM FOR PERCEPTUAL STUDIES

ASR transcription errors highlight speech regions which are problematic with respect to the ASR system's decoding capacities. These speech regions correspond either to intrinsic ambiguities or to some type of variation not properly accounted for in the ASR system's speech model. In any case, from an ASR perspective ASR transcription errors can be viewed as ambiguous speech regions with acoustical and/or contextual confusability. Thus, ambiguities may either arise as a result of a simplified speech model (model bias), or be due to intrinsic spoken language ambiguities (language bias). ASR systems then offer opportunities to imagine innovative tools for the design of perceptual experiments to clarify the role of context in lexical decoding and, with respect to ASR errors, to sort out the respective roles of model and language biases.

The proposed paradigm aims at addressing the role of *context* in disambiguating problematic ASR words *via* perceptual experiments with human transcribers. These experiments are based on stimuli selected from large automatically transcribed speech corpora, with ASR results (presence or not of ASR errors) as control parameters.

The proposed paradigm is an extension of earlier work investigating the perceptual discrimination of frequently misrecognized words such as short grammatical items that possess (near)-homophones [10]. Perceptual experiments in French and American English aimed at identifying 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 results showed that for 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. A clear correlation in lexical transcription success (respectively failure) could be established between ASR systems and humans. The (near)-homophone ambiguity thus 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 an in-depth definition of the *context* parameter.

In the following, we make use of the proposed 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, both for French and English. The target words are frequently misrecognized words, e.g. acoustically poor grammatical words likely to be confounded with corresponding (near-)homophones. For the current experiment these words included *et, est, des, les, à, a* in French. They are observed in various lexical environments according to spoken regions that are erroneously transcribed by the automatic system, i.e., contexts where substitutions, deletions and insertions have been observed. For each target word, embedding stimuli of length 3, 5, 7 and 9 words are extracted from the corpus data, which allows to simulate the maximum language model span of 2-gram, 3-gram, 4-gram and 5-gram language models. Automatic Word Error Rates (WER) for the involved English word pairs are about 15%, whereas they rise above 20% in French broadcast news data [1]. In the present study, experiments were designed to gather more information about the impact of an increasing local context in disambiguating problematic lexical items. The considered 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 earlier fixed 7-gram experiments. Recall that seven word contexts correspond to the maximum span of the ASR 4-gram language model. The proposed experiment makes use of the 2009 QUAERO French and English test data (www.quaero.org).

Table 1. Human WER (in %) for ASR correct and incorrect stimuli according to the n-gram size

WER/n-gram	3-gram	5-gram	7-gram	9-gram
ASR correct	7.3%	1.8%	2.3%	0.9%
ASR incorrect	34%	24%	21%	18%
Global	31%	21%	18%	% 16

Stimuli were selected to contain as a central item one of the target words. Three additional factors *context size*, *automatic transcription of the target word* (i.e. correct vs. erroneous) and *type of automatic error* (i.e. substitution, insertion, deletion) were considered for stimuli selection. A total of 200 target words was selected from the French QUAERO 2009 test data. 10% were correctly transcribed by the ASR system, the remaining target words were either substituted, deleted or inserted. For each of these 200 targets, 4 embedding stimuli of length 3, 5, 7, 9 words were extracted, resulting in a total of eight hundred spoken excerpts. The 800 stimuli were then divided into 4 distinct sets of 200 stimuli, each set including all 200 target words, but with stimuli of randomly varying context size. Each 200 stimuli set was used with a test population of 10 human transcribers. The full test thus required 40 French native transcribers. 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.

Human transcription performance was measured in terms of human WER and compared with the automatic solution for the central targets. Table 1 sums up the human performance in terms of WER on the central target word according to stimulus length (3, 5, 4, 9-gram) and the type of stimulus (i.e. whether or not there was an ASR error on the target word).

The preliminary findings are as follows:

(i) Human WERs decrease with increasing context size: 31% for 3-grams, 21% for 5-grams, 18% for 7-grams and 16% for 9-grams.

(ii) The benefit is particularly high when increasing the context from 3 to 5-gram.

(iii) Automatic and human errors show a positive correlation: results suggest spoken regions erroneously transcribed by ASR system are also challenging for the human transcriptors.

(iv) Target words elicit overall more recognition difficulties: WER averages 22% for each lexical item considered as the center of the n-gram. The result is consistent with previous findings [10].

A sibling experiment for English is presently underway. More extensive studies including a larger range of central targets are foreseen. We will now turn to the description of the graphical interface for ASR error analysis, the Q-Error tool.

4. Q-ERROR GRAPHICAL INTERFACE

This section describes the design and first developments of the Q-ERROR graphical interface for ASR error analysis. Beyond the need for thorough ASR error analyses, the interface is meant to facilitate the selection of controlled stimuli for perceptual experiments of human transcription benchmarks, (such as one we described earlier). Speech transcription errors can be related to various influential factors:

(i) SNR ratio, *(ii)* speaker specificities (e.g., rate, accents, errors, fluency, register), *(iii)* manual reference quality and consistency with ASR (e.g., glm), *(iv)* communication setting (e.g. monologue vs interactive speech) *(v)* automatic processing parameters including the partitioner, the ASR system configuration (e.g., acoustic models, pronunciations, LMs). Among the lexical items affected by one or several of the above-mentioned factors, some segments are more prone to errors than others. (Almost)-homophones, short words, poorly articulated or reduced words, infrequent items in particular proper names, are thus particularly subject to erroneous transcriptions.

The Q-ERROR graphical interface aims both at quantifying and visualizing speech transcription errors according to such taxonomy elements (factors, type of erroneous words, type of errors).

The overall goals of the Q-ERROR tool are the following:

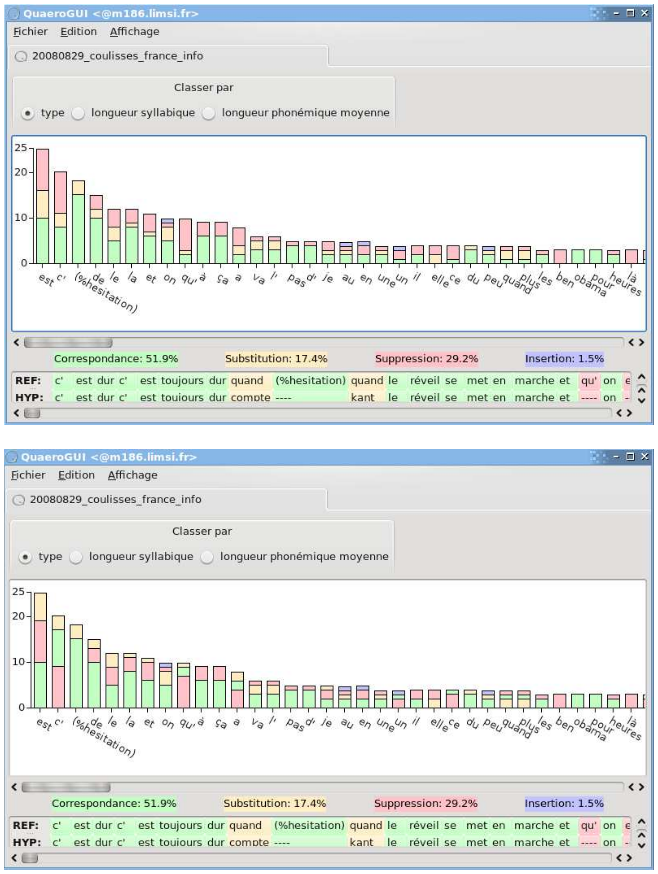
- to selectively look into errors of a given type
- to enable the extraction of typical error samples for perceptual tests
- to improve our understanding of the nature of ASR errors
- to iteratively improve the error taxonomy and related word class definitions
- to propose methods/models to reduce error rates
- to develop a sharable Quaero error analysis tool.

Input information to the Q-ERROR tool comprises reference and hypothesis word strings, corresponding time stamps, pronunciations (either ASR system-specific aligned pronunciations or sharable generic baseform pronunciations), *part of speech* and *named entity* tags as well as system-dependent information (n-gram log likelihood, acoustic confidence scores...). Reference and hypothesis word pairs are checked for temporal overlap and associated a label with respect to automatic transcription: *OK, SUB, DEL, INS* corresponding respectively to *correct, substituted, deleted, or inserted* words. The information may be combined to define various word classes, for which specific error rates can be computed.

The tool should then comply with the following specifications:

1. implement **sets of classes** within the GUI
   (word, word length, frequency, POS, NE tags, speech rate...)
2. produce (correct, substitution, deletion, insertion) rates per **class**
3. listen to audio samples involving specific error classes
4. select samples for perceptual tests

Figure 1 gives a snapshot of the current interface. A first interactive window allows to define word classes. The proposed options are limited to word type, word syllabic length and average phonemic duration. A second window shows the occurrence of each word class in the investigated corpus, with the corresponding proportions of well-recognized items (in green), substitutions (in yellow), deletions (in red) and insertions (in blue). A click on a specific bar section allows to display the corresponding speech segments in a third window (bottom part). These segments may then be listened to, and selected for perceptual tests.


QuaeroGUI <@m186.limsi.fr>
Fichier Edition Affichage
20080829_coulisses_france_info
Classer par
● type ○ longueur syllabique ○ longueur phonémique moyenne
25
20
10
0
est c' (%hesitation) de le la et on qu' à ça a va l' pas d' je au en une un il elle ce du peu quand plus les ben obama pour heures là
QuaeroGUI <@m186.limsi.fr>
Fichier Edition Affichage
20080829_coulisses_france_info
Classer par
● type ○ longueur syllabique ○ longueur phonémique moyenne
25
20
10
0
est c' (%hesitation) de le la et on qu' à ça a va l' pas d' je au en une un il elle ce du peu quand plus les ben obama pour heures là

Correspondance: 51.9% Substitution: 17.4% Suppression: 29.2% Insertion: 1.5%

REF: c' est dur c' est toujours dur quand (%hesitation) quand le réveil se met en marche et qu' on e
HYP: c' est dur c' est toujours dur compte ---- kant le réveil se met en marche et ---- on -

Correspondance: 51.9% Substitution: 17.4% Suppression: 29.2% Insertion: 1.5%

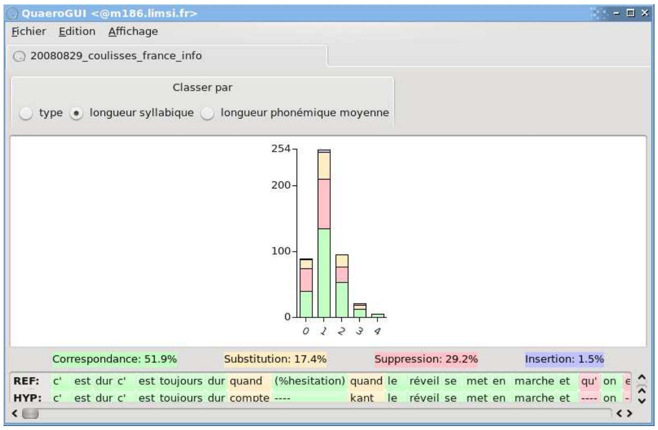
REF: c' est dur c' est toujours dur quand (%hesitation) quand le réveil se met en marche et qu' on e
HYP: c' est dur c' est toujours dur compte ---- kant le réveil se met en marche et ---- on -

Fig. 1. Two snapshots of the Q-Error graphical interface. X-axis sorts word types by decreasing frequency. Y-axis shows error rates, with detailed figures for correctly recognized words (in green) and substitution (in yellow), deletion (in red) and insertion (in blue) error types. In the upper snapshot, classes are stacked using a class order (starting with "correct" and finishing with "insertion"). In the lower part, the 4 (correct + 3 error) classes are stacked according to their frequencies.

Figure 2 gives another screen snapshot, showing the different error types as a function of the number of syllables in the word. Although overall there are about 50% of the words correct, it can be seen that there are more errors on shorter words, in particular with 0 (only a consonant such as *l'*, *n'*, *c'* which are very frequent in French) or 1 syllables. Error distributions are shown as a function of increasing average duration (in seconds) of the phonemes in the word in Figure 3.

5. SUMMARY

This contribution aimed at proposing a paradigm for perceptual experiments to investigate human decoding capacities on ASR error speech stimuli. More specifically, the paradigm aims at assessing 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 word pairs, in French, and evaluated these in a series of perceptual tests. It was



Classer par
○ type ● longueur syllabique ○ longueur phonémique moyenne

Correspondance: 51.9% Substitution: 17.4% Suppression: 29.2% Insertion: 1.5%

```
REF:  c'  est dur c'  est toujours dur quand   (%hesitation) quand le  réveil se  met en  marche et  qu'  on  e
HYP:  c'  est dur c'  est toujours dur compte  ----          kant  le  réveil se  met en  marche et  ---- on  -
```

Fig. 2. Snapshot of the Q-Error graphical interface. X-axis shows word classes of n-syllable length by increasing length in number of syllables. Y-axis shows error rates, with detailed figures for the classes of correct (in green), substituted (in yellow), deleted (in red) and inserted (in blue) words. The classes are stacked according to their frequencies.

shown that human listeners performed 5 to 6 times better than the ASR system on the speech chunks’ central word set. They produced significantly more errors on stimuli misrecognized by the ASR system than on those correctly decoded by the ASR system, where a residual error rate of about 1% was measured. Human transcription accuracy did vary as a function of syntactic and semantic ambiguity. The perceptual tests have been showing that speech errors are typically modulated by a number of factors. In the present paper, we have also reported the development of Q-error, a graphical interface that allows one to filter data in large corpora. By quantifying and visualizing speech transcription, the tool should ultimately enable us to conduct more thorough speech error analyses. The interface can also be used to select the speech stimuli that is needed for further behavioral experiments. Future investigations 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.

6. ACKNOWLEDGMENTS

We express our thanks to Max Pavageau who implemented the first version of the Q-Error graphical interface during a Master summer internship. This work is partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

7. REFERENCES

[1] Adda-Decker, M. and Lamel, L., “Pronunciation variants across system configuration, language and speaking style”, Speech Communication, vol. 29, pp. 83-98 (1999).

[1] Adda-Decker, M., “De la reconnaissance automatique de la parole 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.

[5] Gauvain, J.L. et al., “Where are we in transcribing French broadcast news”, in Proc. of Interspeech, 2005.

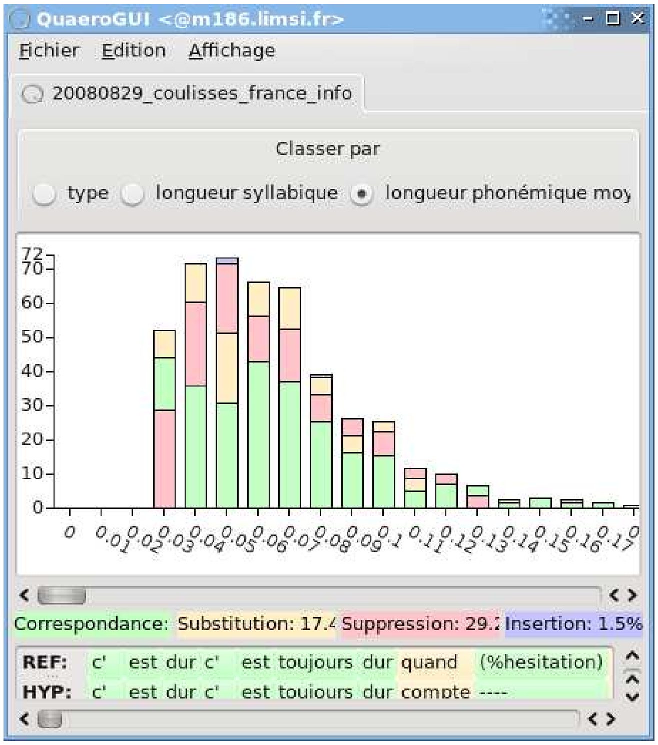
[6] Galliano, S et al., “ESTER PhaseII Evaluation Campaign for Rich Transcription and Broadcast News”, in Proc. of Interspeech, 2005.

[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”, in Proc. of Interspeech, 2008.

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

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



Classer par
○ type ○ longueur syllabique ● longueur phonémique moy

Correspondance: Substitution: 17.4 Suppression: 29.2 Insertion: 1.5%

```
REF:  c'  est dur c'  est toujours dur quand   (%hesitation)
HYP:  c'  est dur c'  est toujours dur compte  ----
```

Fig. 3. Snapshot of the Q-Error graphical interface. X-axis shows word classes of increasing average phonemic length. Y-axis shows error rates, with detailed figures for the classes of correct (in green), substituted (in yellow), deleted (in red) and inserted (in blue) words. The latter classes are stacked according to their frequencies.