ABSTRACT
Automatic speech recognition systems (ASR) have more trouble processing spontaneous speech (e.g. debates) than prepared speech (e.g. broadcast news). These difficulties are due to peculiarities of spontaneous speech (false start, repetition, schwa, etc.). In this paper, we highlight some of these peculiarities, especially in French.

We show that the use of manual transcriptions having no link with the focused application, but which contains only transcriptions of very spontaneous speech, allows to estimate a better language model, strongly decreasing perplexity and significantly decreasing the word error rate on spontaneous speech.

But other knowledge bases used by the ASR have to be adapted. For example, our work shows that adding specific pronunciation variants seems useful, but has to be constrained and modelized. Finally, we compare errors of our CMU Sphinx-based ASR system on spontaneous vs. prepared speech.

General Terms
Automatic speech recognition system, Spontaneous speech, Error analysis.

Keywords
Natural language, French peculiarities, Language model, Pronunciation variants, Broadcast news.

1. INTRODUCTION
Spontaneous speech occurrences in Broadcast News (BN) can take several forms: interviews, debates, dialogues, etc. The main evidences characterizing spontaneous speech are disfluencies (filled pause, repetition, repair and false start) and many studies have focused on the detection and the correction of these disfluencies [8, 7] as pointed out by the recent NIST Rich Transcription Fall 2004 blind evaluation.

All these studies show an important drop in performance between the results obtained on reference transcriptions and those obtained on automatic transcriptions. It can be explained by the noise generated by ASR systems on spontaneous speech segments with higher Word Error Rate (WER) than on prepared speech. Indeed state-of-the-art ASR systems yield high WER when transcribing data likely to contain a lot of spontaneous speech, like conversational speech or meeting recordings. This study tends to highlight the close link between WER and spontaneous speech.

In addition to disfluencies, spontaneous speech is also characterized by ungrammaticality and a language register different from the one that can be found in written texts [1]. Depending on the speaker, the emotional state and the context, the language used can be very different. This context may be particularly conducive to the distortion of some word pronunciations.

This study was made in the context of the EPAC project which focuses on the processing of spontaneous speech. The EPAC project uses some part of the 1700 hours of untranscribed audio files distributed during the ESTER French ASR evaluation campaign. Only audio files containing conversational speech are focused on in this project. Four academic labs are taking part in the project.

In this paper, we will present experiments carried out to show the inadequation of a generic language model (LM) for spontaneous speech, and the effects of its improvement on WER. Then we will focus on analyzing the peculiarities of spontaneous speech, and particularly errors due to pronunciation variants. Finally, we will seek to evaluate the impact of out-of-vocabulary words on the words around them.

2. SUBJECTIVE LEVEL OF SPONTANEITY
By defining spontaneous speech as unprepared speech, we follow a definition proposed by [9] that defined a spontaneous utterance as “a statement conceived and perceived during its utterance”. This definition illustrates the subjec-
tivity of the classification of speech as prepared or spontaneous. Ideally, to annotate a corpus with labels representing the spontaneity of each speech segment, we would have to ask each speaker to annotate his own utterances. This is of course not feasible, however we followed the definition by defining an annotation protocol based on perception by a human judge of a level of spontaneity for a given speech segment. Our approach was to manually tag a corpus of speech segments with a set of ten labels corresponding each to a spontaneity level: grade 1 stands for prepared speech, almost similar to read speech, and grade 10 stands for very disfluent speech, almost non understandable. This approach allowed us to subjectively choose where the limit between spontaneous and prepared speech is placed. In our experiments we considered 3 classes: prepared speech corresponding to grade 1; low spontaneity corresponding to grades 2 to 4; and high spontaneity corresponding to grades 5 and over.

Two human judges separately annotated a speech corpus by listening to the recordings. The corpus was split into segments by a state-of-art automatic segmentation and disambiguation process [2]. No transcriptions were provided to the annotators. Inter-annotator agreement on the spontaneity grades was checked on a separate 1-hour Broadcast News corpus.

2.1 Experimental data

Part of the data used was from the ESTER evaluation campaign [5] on rich transcription of French broadcast news, which took place in 2005. This part included 100 hours of transcribed programs and 1700 hours of non-transcribed recordings of radio broadcasts. The rest of the data came from the EPAC project, which contains more spontaneous speech than the ESTER corpus.

The corpus given to the annotators for labeling was composed of 11 files containing French Broadcast News data from 5 different media (France Culture, France Inter, France Info, Radio Classique, RFI). The files were chosen for being likely to contain spontaneous speech according to the kind of show. The total duration was 11h 37' for 3322 speech segments. 1142 of the segments were labeled as prepared speech, 1175 as low spontaneity, and 1005 as high spontaneity label. No speech segment got annotated with a spontaneity level greater than 7.

2.2 Validation of the subjective annotation method

In order to evaluate inter-annotator agreement for this specific tagging task, we computed the Kappa coefficient of agreement [6] on one hour of BN (distinct from the data described above). The coefficient obtained was very high: 0.852 — a value greater than 0.8 is usually considered as excellent [4].

3. LANGUAGE MODELING

3.1 Generic language model

The training data of the ESTER evaluation campaign was used to train a generic LM. It is composed of manual transcriptions (80h) of broadcast news data, but the main part of the data comes from the French newspaper Le Monde. The amount of very spontaneous speech is limited in this corpus, therefore the main contribution to this LM comes from written texts.

Using the vocabulary (65K words) built for the LIUM participation in the ESTER campaign, the two data sets (Le Monde + BN manual transcriptions) were used to train the baseline trigram and quadriagram LMs. To estimate and interpolate these LMs, the SRILM toolkit [10] was used. Each language model is a backoff model, using the Kneser-Ney modified discounting method with interpolation for low-order n-grams. More details are given in [2].

3.2 LM adaptation

Because the BN corpus presented in section 2.1 contains a lot of spontaneous speech, there is a mismatch between it and the generic LM. We tested an approach based on LM adaptation for dealing with the mismatch. We used an external corpus, which has no link with broadcast news, but which contains transcriptions of very spontaneous speech with explicit annotations of speech disfluencies. The goal was to adapt the generic LM by integrating new information, relevant to spontaneous speech, provided by another knowledge source. The LM adaptation consisted in interpolating linearly the initial generic LM with the new LM trained on the external corpus.

For this, we used manual transcriptions of open conversations made available through project PFC (Phonologie du Français Contemporain).

3.3 PFC corpus

The PFC project [3] involves over thirty researchers from a variety of countries and aims at the recording, partial transcription and analysis of over 500 speakers from the francophone world on the basis of a common protocol. Audio recordings are made mostly of open conversations. This corpus of manual transcriptions of conversations constitutes an interesting knowledge source for modeling spontaneous speech phenomena: we extracted from it 26K sentences with 285K word occurrences in order to build our spontaneous speech LM.

4. ASR SYSTEM DESCRIPTION

Experiments on speech recognition were carried out by using the LIUM ASR system, based on the CMU Sphinx 3.x decoder, described in [2]. It is a three-pass system: the first pass uses a trigram language model and generic acoustic models (one for each of the four gender/band conditions — female/male + studio/telephone), the second pass uses the best hypothesis of the first pass to adapt the acoustic models using SAT and CMLLR, and the last pass consists in rescoring with a quadriagram language model a word-graph generated during the second pass. This system ranked second in the French ASR evaluation campaign ESTER, and was the best open source system.

5. EXPERIMENTS ON LM

In order to evaluate the correlation between WER and the subjective level of spontaneity described in section 2, we transcribed the 11 files described in section 2.1 with the baseline ASR system. Figure 1 shows that there is a real correlation between WER and the subjective level of spontaneity: as soon as a speech segment is not perceived by the human annotators as perfectly uttered (i.e. for a level greater than 1), the WER increases.
The 11 files were divided into two corpora: a development corpus (7 files) and a test corpus (4 files, about 4h 15'). Table 1 shows the distribution of the three spontaneity classes in the test corpus.

<table>
<thead>
<tr>
<th>Subjective level of spontaneity</th>
<th>Prepared</th>
<th>Low Sp.</th>
<th>High Sp.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># segments</td>
<td>1314</td>
<td>1132</td>
<td>1715</td>
<td>4181</td>
</tr>
<tr>
<td># words</td>
<td>13493</td>
<td>12218</td>
<td>19292</td>
<td>45008</td>
</tr>
</tbody>
</table>

Table 1: Number of speech segments and words in the text corpus for the three classes of spontaneity

The development corpus was used to optimize the interpolation coefficient between the generic LM and the spontaneous speech LM. We refer to the interpolated LM as Base+PFC.

Table 2 shows the results in terms of perplexity on the test corpus according to the spontaneity class. As we can see, using a LM specific to spontaneous speech significantly reduces perplexity for the classes low spontaneity and high spontaneity while, as expected, perplexity increases for the class prepared speech. It is interesting to note that the PFC corpus brings a notable gain (-19.21%) only on the high spontaneity class. This can be explained by the thematic mismatch between this corpus and broadcast news data. Highly disfluent speech though can benefit from this corpus despite the mismatch.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>21.4%</td>
<td>31.5%</td>
<td>41.2%</td>
<td>32.6%</td>
</tr>
<tr>
<td>Base+PFC</td>
<td>21.7%</td>
<td>31.5%</td>
<td>40.1%</td>
<td>32.4%</td>
</tr>
</tbody>
</table>

Table 2: Perplexity for each LM and spontaneity class

WER results are presented in table 3. The values are relatively high because the files chosen for the experiments contained much more spontaneous speech than the ones used during the ESTER evaluation campaign.

In the same way as for the perplexity results, we can notice that using the adapted LMs reduces WER on the highly spontaneous speech while it increases the WER on other kinds of speech. The improvement is statistically significant on highly spontaneous speech according to the 95% confidence interval criterion. However, when compared to the observed perplexity decrease on highly spontaneous speech, the WER decrease is disappointing: other resources than language models have to be taken into account and modeled to improve ASR processing on spontaneous speech. Word pronunciation will be investigated below.

<table>
<thead>
<tr>
<th>Word error rate</th>
<th>Subjective level of spontaneity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 1: Word error rate according to subjective spontaneity level

6. PRONUNCIATION DICTIONARY

6.1 Pronunciation rules

Pronunciations used in the dictionary of an ASR system developed for broadcast news do not really seem to suit spontaneous speech. Missing pronunciations generate deletions and substitutions. We wanted to evaluate the effect of adding some specific pronunciations variants for the concerned words.

For this study, we chose to focus on one particular problem of spontaneous speech in French, concerning the vibrant phone /r/. Especially for words ending in “-bre”, “-cre”, “-dre”, “-tre” or “-vre”, uttering of the /r/ is mandatory in a correct language style. For example, the verb “ouvre” has to be pronounced /uvr/. In spontaneous speech however, the /r/ is sometimes omitted at the end of a word even while it is incorrect, and thus the pronunciation /uv/ can occur.

We added such pronunciations for this kind of words. Furthermore, pronunciations specific to spontaneous speech were added for a list of words frequently used in French (“pour”, “elle”, “enfin”, “il”, “je”, and “parce que”). Initially, the dictionary contained about 165K pronunciation variants for 62K words. 1761 variants were added, affecting 638 words (1.02% of the words).

6.2 Experiment

Some part of audio files of the EPAC project (14 hours in total) with a very high proportion of spontaneous speech were used to evaluate the impact of adding special pronunciations.

Out of the 13K distinct words present in these files, 203 words (1.52%) were words for which we had added pronunciation variants. This is pretty similar to the proportion cited above of words with new variants in the dictionary.

When it comes to occurrences, however, the proportion is quite different: 8.35% of the 178960 occurrences of words in these files are words with new pronunciation variants. This high rate highlights the fact that we added pronunciation variants to some words very widely used in oral French.

6.3 Results

Table 4 presents the effect of the addition of pronunciation variants on the substitution and deletion errors for the concerned words.

As expected, the new pronunciations had a positive impact on the substitution and deletion mistakes of the affected words. It is also interesting to know how the new pronunciations allowed a better decoding of these words. 500 occurrences of words were decoded using these pronunciations: it
Table 4: Error (substitution + deletion) count and rate for the words concerned by added pronunciation variants

<table>
<thead>
<tr>
<th>Count</th>
<th>Base</th>
<th>Base + pron. variants</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>22.54%</td>
<td>20.82%</td>
<td>-1.72%</td>
</tr>
</tbody>
</table>

Table 5: Global error (substitution + deletion) counts and rates before and after adding pronunciation variants

<table>
<thead>
<tr>
<th>Count</th>
<th>Base</th>
<th>Base + pron. variants</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>25.99%</td>
<td>25.18%</td>
<td>-0.81%</td>
</tr>
</tbody>
</table>

Table 6: Count and rate of word occurrences correctly recognized whereas they are not associated in the ASR dictionary with the pronunciation chosen during the alignment process

<table>
<thead>
<tr>
<th>Correctness rate</th>
<th>Spontan. (EPAC)</th>
<th>Prepared (ESTER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>49.51%</td>
<td>71.85%</td>
</tr>
</tbody>
</table>

Figure 2: Word error rate around an error (all the errors vs. errors due to OOV words) on spontaneous speech and prepared speech

7. ERROR ANALYSIS

At last, for the two entire corpora, we looked around each misrecognized word to analyze error propagation in the neighborhood of this word. This neighborhood is characterized by its a radius n: this corresponds to the n words to the left of the misrecognized word and the n words to the right. We computed the WER on the words for values of n varying from 2 to 5. We distinguished two cases: when the misrecognized word is an out-of-vocabulary (OOV) word, and the general case. Figure 2 shows the corresponding error rates (from n = 2 to n = 5).

This result shows that the WER around a misrecognized word has a tendency, in a general way, to decrease when the neighborhood size increases. On the other way, the WER tends to increase when we focus on the neighborhood of errors due to an OOV word. The same observation can be made on both corpora (EPAC and ESTER).

8. CONCLUSIONS

In this article, we highlighted the need of a specific treatment of spontaneous speech. We integrated a non-thematic corpus containing manual transcriptions of conversations into the estimation of an adapted LM. This LM allowed WER to decrease significantly for highly spontaneous speech, but improving the LM does not appear as the only solution to strongly improve recognition of conversational speech.

Indeed, classical pronunciation dictionaries used in broadcast news transcription seem to be limited for processing spontaneous speech, which requires specific features, such
as to model assimilations. Adding specific pronunciation variants to these dictionaries leads to a reduction in deletions and substitutions, but have to be constrained to limit insertions.

The results are encouraging, although it should be pointed out that our tests concerned a limited number of pronunciation rules. Moreover, we showed that the ASR system has more difficulties to compensate the imprecision of the dictionary of pronunciations when it is used on spontaneous speech. This enforces the idea that word pronunciation in spontaneous speech must be studied in detail in order to obtain better results. Finally, we showed that OOV words lead to a high propagation of errors in their neighborhood, for spontaneous speech as well as for prepared speech. This result is interesting because it is not the same in the neighborhood of other kinds of errors.

This work constitutes a comparative study of the use of an ASR system on prepared speech and on spontaneous speech. In the future we will explore the different paths highlighted in this paper, such as LM adaptation or word pronunciation.

9. ACKNOWLEDGMENTS

The author would like to thank Pr. Paul Deléglise and Dr. Yannick Estève for their supervision work, and Vincent Jousse and Thierry Bazillon for their participation in the tagging with spontaneity level, presented in section 2 Subjective level of spontaneity.

10. REFERENCES


