French Broadcast News Transcription

Robert Hecht

SAIL LABS Technology
Operngasse 20 B, 1040 Vienna, Austria
robert.hecht@sail-technology.com

Abstract

We describe a French broadcast news transcription system created in the scope of the CIMWOS[1] project [1]. We collected a corpus based on two French and one Belgian TV stations. This corpus forms the base of various system components, such as ASR and Speaker ID. We discuss a few problems posed to speech recognition by characteristics of the French language and approaches to solve them. Finally we analyze the performance of the system and its weaknesses and show that a considerable part of the problems is caused by the abundance of homophones in French.

1. Introduction

French orthography is very conservative and reflects many etymological and grammatical phenomena that are no longer of relevance in contempory spoken French. This causes a high number of homophones (words with different spelling but identical pronunciation).

Among the problems that French poses for ASR are:

- Homophonic grammatical forms:
  - Participles: male and female forms of past participles often are homophones, also singular and plural (except in the context of *liaison*, see below). E.g. *allé / allée / allés / allées* (past participle of to go - male and female form, singular and plural). Also, singular and plural of the present participles are often homophones. However, these are not so important because they are less frequent.
  - First, second, and third person singular and third person plural have the same pronunciation for most verbs in most tenses, e.g. *reste / restes / reste / restent* (I stay / you stay / he (she, it) stays / they stay). Note: depending on tense and conjugation, first and second or first and third person singular also coincide in spelling.
  - Besides this "systematic" homophony, sometimes quite unrelated verb forms have the same pronunciation, e. g. *rester / resté* (to stay - infinitive / participle).
- Sometimes the trigrams commonly used for language modeling cannot resolve these ambiguities, because dependencies stretch over more than three words. E.g. in *Il n'est pas encore arrivé* (He hasn't come yet) the pronoun *Il* which determines that the verb form is *arrivé* (male) rather than *arrivée* (female) is more than two words away. Even worse, in sentences like *Je me suis levé* (I got up), the ambiguity (*levé* vs. *levée*) cannot be resolved by any language model, but only by knowing if the speaker is male or female. However, this could be viewed as a scoring issue.
- Besides different grammatical forms of one word with the same pronunciation, there is also a considerable number of unrelated words which have homophonic pronunciations at least for some of their forms, e. g. *nez* (nose) and *né(e)* (born).
- The *liaison*: trailing consonants are pronounced or suppressed depending on whether the next word starts with a vowel: e. g. *les* is pronounced $[le]$ in *les femmes* but $[lez]$ in *les amis*. Other consonants that are normally not pronounced except with liaison are $[t]$ (written as *t* or *d*), $[p]$, $[n]$ (after nasal vowels) and $[R]$. This has the effect that some grammatical forms coincide only if the next word does not start with a vowel e. g. *état* and *états* are both pronounced $[eta]$ if followed by consonant, but the latter changes to $[etaz]$ if followed by a vowel. To complicate things further, liaison is sometimes ignored in casual speech.

These phenomena lead to an unusually high ratio of possible spellings per pronunciation: 1.51 for our decoding vocabulary with 50k Words. Whereas English spelling is also largely determined by etymology, different grammatical forms of a word are pronounced differently (with almost no exceptions). Moreover, in English same pronunciation generally implies same spelling, which is reflected in the fact that for our English system [6] we find only 1.04 spellings per pronunciation (and almost exactly the same value for our German system [7]). We therefore expect a somewhat higher Word Error Rate (WER) for French.

2. Corpora

2.1. Acoustic Model

Our training (test) set consists of 122 (5) TV news broadcasts from two French (TV5, La Cinquième) and one Belgian (RTBF) channel, recorded between October 2001 and March 2002 (see Table 1). Transcription was done using the Transcriber tool [2] whose output is a textual version of the transcripts with markup for speaker turns, speaker names, genders, and non-speech events.

2.2. Language Model

The Language Model (LM) consists of the audio transcripts (804k Words) and the editions of the newspaper 'Le Monde' from 1996 through 2000 (115M Words), available from ELRA.

[1] **C**ombined **IM**age and **WO**rd **S**potting – funded by the Information Society Technologies Programme under contract: IST-1999-12203

Table 1: *Overview of collected audio data in $hh:mm$ per channel. As opposed to the net data the raw data set may also contain (non-speech) sections that were not employed for training of acoustic models.*

channel	**training**		**test**	
	raw	*net*	*raw*	*net*
RTBF	28:17	24:47	1:25	1:31
TV5	27:53	21:25	0:28	0:23
La Cinquième	1:50	1:22	0:55	0:44
total	58:00	47:35	2:48	2:19

In the LM training, the audio transcripts were emphasized by a weighting factor.

We adapted a two-level approach to prepare the corpus for the LM training: in the first step, we converted the input texts to XML with markup for all entities that have to be processed for tokenization, e.g. numbers, dates, abbreviations, interpunction, case naturalization, etc. These XML files are then converted to the format needed by the training programs in a second step. This has the advantage that the XML can be inspected and manually corrected and errors in the tokenization can more easily be traced. An example of the XML used is given below.

```
<Phrase>
<CaseNatural replacement="deux" word="DEUX"/>
<CaseNatural replacement="mètres"
word="MÈTRES"/>
<Interpunction canEndPhrase="0" value=","/>
<Cardinal original="150" value="cent cin-
quante"/>
kilos
<Interpunction canEndPhrase="1" value=":"/>
Marion
<Interpunction canEndPhrase="0"
value="&amp;quot;"/>
<OOV word="Suge"/>
<Interpunction canEndPhrase="0"
value="&amp;quot;"/>
<OOV word="Knight"/>
en impose même au plus retors
<Interpunction canEndPhrase="1" value="."/>
</Phrase>
```

Enclitics (eg. the *-moi* in *donne-moi* - give me) were treated as separate words. This is justifiable because treating words like *donne-moi* as one would inflate the dictionary and the enclitics are long enough to be detected in the audio stream. With proclitics (e. g. *l'* in *l'ami* - the friend) this was less obvious because those consist only of just one consonant and therefore are less likely to be recognized as separate words. Therefore we tried both to split them off and to leave them with the following word. The experiments show that this affects the coverage of the test set with a given lexicon size, but has very little effect on the Word Error Rate (WER).

3. Experiments

3.1. ASR system

A speaker-independent, large-vocabulary speech recognizer based on BBN's Rough'n'Ready suite of technologies [3, 4, 5] is used in our experiments. It employs continuous density Hidden Markov Models (HMM) for acoustic modeling and word-form n-grams for language modeling. The recognition process is implemented as follows: at first a cepstral analysis of the acoustic waveforms is performed to extract a feature vector for each frame of speech. These feature vectors serve as input for three decoder passes. The first (fast-match) pass uses a phonetically tied-mixture (PTM) model based on within-word triphones and a bigram language model. The second (backward) pass employs a state-clustered tied-mixture (SCTM) model depending on within-word quinphones and a trigram language model to produce a set of N-best hypotheses. Then a third, more detailed analysis using a cross-word SCTM model is used to rescore the N-best list to select the final answer.

3.2. Phone-Alphabet and Pronunciation Dictionary

Our Phone-Alphabet consisted of 38 phones for speech (18 vowels, 20 consonants) and 4 for non-speech events: (SILence, BReaTh, LIPsmack, and GaRBage). The Pronunciation Dictionary was based on BDLEX-50000, available from ELRA. However, this dictionary contains phone symbols that can be pronounced in more than one way, (to ensure that there is only one pronunciation per entry), and we mapped them to (different) unique symbols. Where the dictionary indicated the possibility of *liaison* (i. e. the pronunciation of a word depends on whether the next word starts with a vowel or consonant), we put both pronunciations in the dictionary. The experiments indicate that no further handling of liaison is necessary.

The recognition vocabulary was formed from the word of the audio corpus and the most frequent words of the language model. Pronunciations for words that were not in the original dictionary were generated using a grapheme-to-phoneme conversion tool. We experimented with different vocabulary sizes measuring test set coverage and WER.

3.3. Results

3.3.1. OOV Rates

As described in section 3.2, we studied systems with the proclitics as they appear in normal texts (plain) or separated from the following word. Figure 1 shows the Out-Of-Vocabulary (OOV) Rates for both approaches. It can be seen that the former approach leads to higher OOV Rates than the latter, even at large vocabulary sizes. This was to be expected, since one unknown word in the latter approach can produce more than one OOV in the former if it starts with a vowel (*xxx, d'xxx, l'xxx* etc. are counted as separate forms).

The OOV rates are higher than for comparable english systems, but lower than for German [6, 7].

3.3.2. Word Error Rates

Figure 2 shows the WER for both the simple system and the one with separated proclitics. Separating the proclitics gives better WERs at vocabulary sized up to 50k words; for larger vocabularies both systems give almost equal performance. Besides, inflating the vocabulary beyond 70k words does not significantly reduce the WER, which seems logical, because a) further increase causes only very little decrease of the OOV rate and b) at

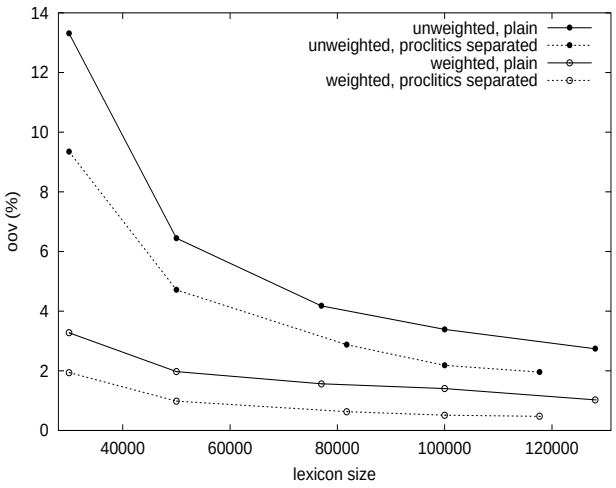

Figure 1: OOV rates versus vocabulary size for the plain system and the one with separated proclitcs. Unweighted percentages take only unique words into account whereas the weighted numbers are calculated using total word counts.

this vocabulary size the OOV errors are just a small percentage of the total errors[2] and therefore reducing the OOV rate does not help.

Since the system with separated proclitics gives the better WER, we base all remaining experiments on it.

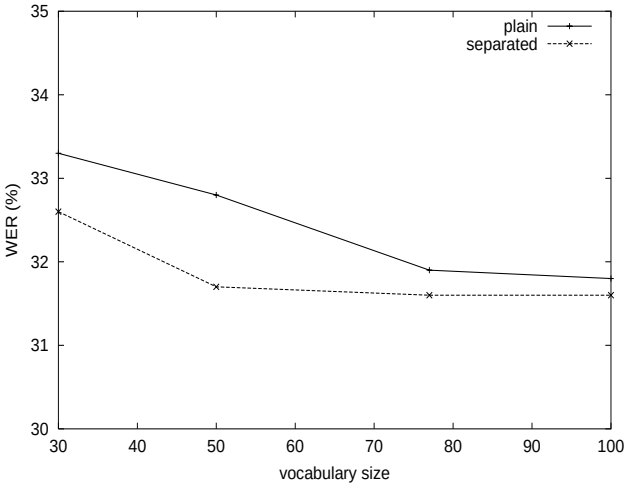

Figure 2: *Word Error Rates vs Vocabulary size.*

Breaking down the WER for the individual channels reveals that the test set is rather inhomogeneous in this respect (see Table 2). The WER for La Cinquième is almost twice as high as that for RTBF. This is because the former contains a high percentage of spontaneous speech (interviews, discussions), whereas the latter consists primarily of read speech. Since our language model is primarily based on newspapers and contains only very little material from La Cinquième, it is clear that it is less suited for spontaneous speech. Besides, disfluencies, dialects etc. make discussions hard to transcribe.

[2] A rule of thumb says that one OOV causes on average 1.5 errors.

The last column of table 2 shows the number of non-speech events per hour, which is a good measure for the percentage of spontaneous speech.

Table 2: *Word Error Rates depending on Channel.*

channel	**WER**			**Nonspeech**
	30k	50k	77k	
RTBF	24.0	23.1	23.1	131
TV5	33.6	32.1	32.2	183
La Cinquième	44.1	43.4	43.2	514
total	32.6	31.7	31.6	276

3.3.3. Detailed Analysis of the errors

An analysis of the so-called *confusion pairs* (i.e. which word was substitued by which other word in the recognition process) and their relative frequencies reflects the problems listed in section 1. Specifically, we examined the following subgroups of confusion pairs with equal pronunciation (the labels at the begin of each line identify the entry in the table):

- NumNoun: singular vs. plural of nouns and adjectives (e.g. *raison* vs. *raisons*)
- Part: different forms of past participle (e.g. *arrivé* vs. *arrivée*)
- VerbFinite: singular vs. plural in finite verb forms (e.g. *arrive* vs. *arrivent*)
- VerbOther: other mixups of verb forms (e. g. *arriver* vs. *arrivé*)
- BasePron: homophones derived from the same base form.
- Base: all pairs with a common base form, but possibly different pronunciations (e.g. different forms of a verb)
- Pron: all homophones (possibly stemming from different words)

Table 3 shows that up to 20 % of the total WER can be caused by homophones. The phenomena that contribute most to this kind of error are confusion of singular vs. plural of nouns and adjectives and forms of past participles. Although the percentage of homophony-induced errors is less in the La Cinquième test set, the table shows that the contributions of the listed subclasses remain almost constant accross test sets, as can be seen by dividing all figures by the contents of the row ”Pron”. This indicates that there might be a lower limit on the WER induced by homophones alone. But this may be true only in the context of trigram LMs; other types of LMs taking into account grammatical dependencies and sentence structure might be able to overcome this limit.

4. Conclusions

We described a newly established corpus of French broadcast news data and the development of an ASR system based on this corpus. Challenges posed by the French language have been described and their effects on the system performance have been investigated. Even though we use a very simple word-based trigram model, the overall performance is satisfactory. Detailed analysis of the errors commited confirms the anticipated problems and indicates areas for further improvement.

Table 3: *Frequency of different types of confusion pairs (percentage of total WER.*

Type	RTBF	TV5	La Cinquième
NumNoun	5.5	4.1	3.0
Part	3.9	4.8	1.4
VerbFin	1.3	1.7	0.4
VerbOther	1.0	1.2	0.7
BasePron	12.4	13.7	6.3
Base	14.5	14.4	7.7
Pron	19.1	18.3	10.3

5. Acknowledgements

Thanks to Jürgen Riedler and Gerhard Backfried for fruitful discussions and to Sabine Loots for correcting my French.

6. References

[1] ⟨http://www.xanthi.ilsp.gr/cimwos⟩

[2] Barras, C., Geoffrois, E., Wu, Z., and Liberman, M., "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech", Proceedings of the 1st International Conference on LREC, 1998.

[3] Kubala, F., Jin, H., Nguyen, L., Schwartz, R., and Matsoukas, S., "Broadcast News Transcription", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'97 (Munich, Germany), 1997.

[4] Nguyen, L., Matsoukas, S., Davenport, J., Liu, D, Billa, J., Kubala, F., and Makhoul, J., "Further Advances in Transcription of Broadcast News", Proceedings of the 6th European Conference on Speech Communication and Technology, EuroSpeech'99 (Budapest, Hungary), 1999.

[5] Colthurst, T., Kimball, O., Richurdson, F., Han, S., Wooters, C., Iyer, R., and Gish, H., "The 2000 BBN Byblos LVCSR System", Proceedings of the 6th International Conference of Spoken Language Processing, ICSLP'2000 (Beijing, China), 2000.

[7] Hecht, R., Riedler, J., Backfried, G., "German Broadcast News Transcription", ICSLP 2002, Denver, Colorado

[6] Backfried, G., Hecht, R., Loots, S., Pfannerer, N., Riedler, J., Schiefer, C., "Creating a European English Broadcast News Transcription Corpus and System", Eurospeech 2001, Aalborg, Denmark.