Experiments on the accuracy of phone models and liaison processing in a French broadcast news transcription system

Dominique Fohr, Odile Mella, Irina Illina, Christophe Cerisara

Speech Group LORIA-INRIA, France
www.loria.fr/Equipe/parole

Abstract

In the framework of ESTER, the recent French broadcast radio news transcription task evaluation, we have developed the first version of ANTS, the Automatic News Transcription System of LORIA. This paper describes the different components of the ANTS system and provides some first recognition results on the ESTER database. Then it presents several experiments carried out on this system to take into account the specificities of the French language: how accurate should the phones models be and how to deal with the problem of the liaisons between words.

1. Introduction

The transcription of broadcast radio news poses a number of challenges for large vocabulary recognizers. The transcription system has to deal with a continuous audio stream containing various types of speech (spontaneous, quickly read and conversational), uttered by native or non-native speakers in different environments (music/noise background, telephone channel, outside conditions). All these factors should be taken into account by the transcription system.

Several transcription systems have been proposed by different research laboratories: HTK system [1], BBN system, LIMSI system [2] in the framework of several DARPA and NIST evaluations. Since 2003, the French Department of Research has organized a similar evaluation campaign for the French language: ESTER (Evaluation des Systèmes de Transcriptions Enrichies des émissions de Radio) [3]. Therefore, we have developed a first version of ANTS: the Automatic French News Transcription System. During this recent development we have carried out some experiments about the specificities of the French Language. We have first studied the accuracy of the phone models used by the recognition engine to take into account the phenomenon of the archiphonemes. We have also investigated the best manner to deal with the problem of the liaisons between words.

This paper is organized as follows: section 2 briefly describes the database; section 3 presents the different components of the ANTS system as well as some recognition results achieved on the ESTER database; section 4 studies the influence of the accuracy of the French phone models and section 5 the various manners to process the liaison between words.

2. Database description

2.1. Corpus

The training and test data were extracted from the Radio broadcast news corpus created in the framework of the ESTER project. While our developments were in progress, only the first part of the ester corpus was available. This part contains 40 hours of manually transcribed shows from two French speaking radio stations (France Inter and Radio France International). This corpus is divided into three parts: 30h for the training, 4h40 for development and 4h40 for test.

We have conducted our experiments on the whole test corpus, but we have only used 7 hours of the training corpus. Indeed, we have removed:

- sentences containing proper names for which we do not have the phonetic transcription,
- narrow band (telephone) speech,
- music segments,
- speech segments with music in background.

The remaining segments have been used for training HMM phones models for broadband speech.

Narrow-band HMM models have been trained on the telephone corpus SPEECHDAT1000 (read speech by 1000 speakers).

3. ANTS: Automatic News Transcription System

As shown in Figure 1, the Automatic News Transcription System (ANTS) developed at LORIA is composed of four stages: broad-band/narrow-band speech segmentation, speech/music classification, detection of silences and breath noises and large vocabulary speech recognition. The aims of the three first stages are to split the audio stream into homogeneous segments with a manageable size and to allow the use of specific algorithms or models according to the nature of the segment. The following sections describe these four stages.

3.1. Parameterization

The speech signal is sampled at 16 kHz. For analysis, 32 ms frames are used, with a frame shift of 10 ms. 13 MFCC are calculated for each frame and the 13 delta and 13 delta-delta coefficients are added, leading to a 39 dimension vector. In all experiments presented in this paper we use CMR (Cesptral Mean Removal).

3.2. Broad-band/narrow-band speech segmentation

This segmentation module labels audio frames according to bandwidth. It is based on the ratio of energy above 4kHz to that between 300Hz and 4kHz. This ratio should be small when the speaker uses the telephone network.

This module outputs segments labeled as broadband speech or narrowband speech.
3.3. Speech / Music classification

The topic of this stage is to remove the non-speech segments from the audio stream. It is only applied on the broadband tagged segments because we assume there is no music during a phone call. Our classifier is based on four competing models for speech, instrumental music, songs and speech with background music. Each model consists in a Gaussian mixture model (GMM) with 32 pdfs and diagonal covariance matrices. As the duration of a segment of speech or music signal is usually greater than a second, even for jingles, we impose a minimum duration for each recognized segment. This is done by duplicating 100 times the state of each GMM as shown in Figure 2.

During the segmentation phase of an unknown signal, the Viterbi algorithm computes the best sequence of the competing models. All the segments labelled as instrumental music or songs are removed.

3.4. Breath noises and silences detection

The aims of this step are first to split the audio stream into smaller segments than can be processed by the recognition engine and secondly to provide segments related to the prosodic phrasing. This detection of breath noises and silence phases is performed by a phone recognizer. It uses 36 context independent phone models, a silence model and a breath noise model with a null language model. This detection is applied on the narrow-band and broad-band segments.

3.5. Large vocabulary speech recognizer

The large vocabulary speech recognizer is based on the Julius recognition engine originally developed by Akinobu Lee at Kyoto university [4]. Speech recognition is performed in two passes: in the first pass, a tree-structured lexicon assigned with bigram language model probabilities is applied with the frame-synchronous beam search algorithm. This pass gives a word-trellis. In the second pass a stack decoding algorithm with trigram language model gives the N-best recognition sentences. Julius can be compiled into 3 versions: fast, standard and v2.1. Fast version allows high speed but the precision decreases compared to standard and v2.1 versions. The v2.1 is the slowest but is the most accurate with bigram factoring and word-pair approximation.

For this broadcast news transcription task, we trained two sets of 38 context-independent phone models including one silence model and one breath model. One set of such models are trained for telephone speech and another for broadband speech. The models are HMM with 3 states, 256 gaussian mixtures per state and diagonal covariance matrices. The language model was trained with the CMU toolkit, resulting in 2.5 M bigrams and 5.8 M trigrams. We built a 54547-words lexicon composed of the most frequent words extracted from the 1995-2001 years of the French newspaper “Le Monde” plus all the words occurring at least 3 times in the orthographic transcriptions of the ESTER training corpus. The common nouns were phonetized with the French Lexicon BDLEX [5] but the proper ones were manually phonetized.

3.6. First recognition results

We evaluated this first version of our transcription system on the ESTER dry run corpus with the following parameters:

- 38 context-independent HMM models with 256 pdf: 36 phones models plus one model for silence and another for breath noises. The covariance matrices are assumed to be diagonal.
- a lexicon containing 54747 words and 58588 phonetic pronunciations,
- a language model: 2.5 M bigrams, 5.8 M trigrams,
- Adaptation: MLLR, block diagonal ,
- Julius: v2.1 version (highest precision),
- a beam set to 2000.

Recognition results are presented in Table 1 as word error rate averaged over all test speakers.
2, we have decided to test five phonetic lexicons. Various phonetic transcriptions for a word. As shown in Table 1 described. Indeed, given the set of models, we can choose the phase depends on the three sets of models previously described. Besides, the phonetic lexicon used during the recognition process is neutralized in some words. Therefore, the speaker can utter one or the other phoneme in these words, or a phone acoustically situated between the acoustic realizations of the two phonemes. For instance, if we consider the couple (ε, e), the word “chantait” (he sang) must be pronounced with the open vowel /ε/ and the word “chanter” (to sing) with the close vowel /e/, but the word “maison” (house) can be pronounced with an /ε/ or an /e/ or any allophone between these two phones. Among the French vowels, this phenomenon can be also observed with (O, o) and (œ, œ) [6].

Because of this, the degree of accuracy in the phones’ definitions must be chosen when designing a recognizer. The question is: shall the recognition engine use only one model for the archiphoneme or two models or three? We have decided to address this problem in the framework of our automatic transcription system. However, we have limited our experiments to the (ε, e) pair because the tests are very time consuming.

We trained three model sets named “1ε”, “2ε” and “3ε” respectively containing:
- only one rough model for /ε/, /e/ and their allophones,
- two models: one for /ε/ and one for /e/. Before training the models, for each occurrence of a word containing the archiphoneme, we have to label this archiphoneme with the open or the close vowel,
- three models: one for /ε/, one for /e/ and another, named /E/, for the occurrences of the archiphoneme in neutralized localizations, as in “maison”.

Besides, the phonetic lexicon used during the recognition phase depends on the three sets of models previously described. Indeed, given the set of models, we can choose various phonetic transcriptions for a word. As shown in Table 2, we have decided to test five phonetic lexicons.

<table>
<thead>
<tr>
<th>1 phonetization</th>
<th>2 phonetizations</th>
<th>3 phonetizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 e m E z ō</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 e m e z ō</td>
<td>m e z ō</td>
<td></td>
</tr>
<tr>
<td>3 e m E z ō</td>
<td>m e z ō</td>
<td>m e z ō</td>
</tr>
</tbody>
</table>

Table 2: Phonetic transcriptions for “maison” related to the phone sets.

To test these five solutions, we have used our ANTS system with the following settings: context-independent phone models with 256 pdfs, a 54747-word lexicon and the standard version of Julius. We explored different values of the beam width to take into account the variation of the size of the phonetic lexicon due to the variable number of phonetizations of a word. The phone sets were trained on the 7 h training corpus and we used a one hour-long file of France-Inter broadcast news for the test.

Figure 3 presents the Word Error Rate as a function of recognition time, which itself depends on the beam width. We can observe that it is better to train three different accurate phone models providing that you use only the third model (/E/) in the phonetic transcription of the words where the distinction between (ε, e) is neutralized as in “maison”.

We have conducted this study only for the couple (ε,e) and (O,o) and for context-independent models. Therefore, we will have to confirm this conclusion, on the one hand by testing the two other couples (O, o) and (œ, œ), and on the other hand by performing this study on context-dependent models, although we believe that these conclusions should remain true for triphones.

We can notice that in the recognition test conducted for the dry run evaluation ESTER we used the “3ε_4phon” solution; that is three phone models for each pair: (ε, e) (O, o) and (œ, œ) and only one phonetic transcription for every word.

5. Influence of liaison in French

French liaison consists in producing a normally mute consonant before a word starting with a vowel or an “h”. For instance, the word “les” (the in the plural) is pronounced /l e/ if the following word begins with a consonant and /l e z/ if the following word begins with a vowel.

Depending on the grammatical type of a word - and, sometimes on the grammatical type of the following word - the liaison can be optional, mandatory or forbidden [6][7].

The aim of this study is to evaluate various manners to deal with the problem of French liaison. The liaisons can be taken into account at the lexicon level, at the language model level or in the acoustic models. We have chosen to work at the lexicon level because it is easy to implement and it allows us to do these preliminary tests regardless of the language model. Table 3 shows the Word Error Rate for the following liaison representations:
- “No liaison” means that the liaison is always forbidden,
- “Forced liaison” means that liaison is assumed to be always pronounced even if the following word begins with a consonant. It is obviously a bad idea, but it is also always pronounced even if the following word begins with a consonant.

![Figure 3: WER according to the phone set and the lexicon](image-url)
interesting to measure the resulting decrease in recognition accuracy.

- “4 transparent words”: we added 4 words (/t/, /z/, /R/, /n/) corresponding to the potential liaison phonemes in French. The recognizer can insert one of these words if it “assumes” that a liaison phoneme has been pronounced. But it does not take into account this word when it applies the language model. The drawback of this very simple method is that the recognizer can insert a liaison phoneme even if the word does not end with this consonant.

- “Addition of 80 words with liaison”: for 80 frequent words, we explicitly modelled the liaisons. For each of these 80 words, we added to the lexicon the pronunciation including the liaison phoneme: “les: /l e z/”. These 80 words have been chosen because they belong to a grammatical class for which the liaison is mandatory, such as determiner, monosyllabic adverb, pronoun, monosyllabic preposition and auxiliary verbs at the 3rd person.

- “Addition of 21497 words with liaison”: we added a pronunciation with the liaison phoneme for all the lexicon words where a liaison can occur, namely, which can produce a liaison with a following word.

- “Liaison with a skip transition”: to avoid the duplication of words with liaison, we added a skip transition from the last HMM state of the last model of the word without liaison to the final state of the phonetic transcription (cf. Figure 4). This method results in the addition of 21497 skip transitions in the lexical tree.

For the skip-transition option, we have also tuned the probability to skip the liaison. The various tested values and the corresponding Word Error Rate are presented in Table 4.

<table>
<thead>
<tr>
<th>Skip Probability</th>
<th>WER %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>35.9</td>
</tr>
<tr>
<td>0.1</td>
<td>35.2</td>
</tr>
<tr>
<td>0.5</td>
<td>35.2</td>
</tr>
<tr>
<td>0.9</td>
<td>35.0</td>
</tr>
<tr>
<td>0.99</td>
<td>35.3</td>
</tr>
</tbody>
</table>

Table 4: Word Error Rate in function of the probability to skip the liaison phone.

6. Conclusion

In this paper, we have presented the first version of ANTS, our French Automatic News Transcription System. During the design of this system we focused our work on some specificities of the French language as the archiphonemes and the liaisons. Regarding the former topic, we have shown that it is better to train three phone models for the archiphonemes \((\varepsilon,e)\), providing that you use only the third model (/E/) in the phonetic transcription of the words where the distinction between \((\varepsilon,e)\) is neutralized. As for liaisons, we evaluated different manners to take them into account. Implementing the liaison with a skip transition is the best solution for accuracy as well as for computation time. However, we are testing another solution which consists to integrate the liaison in the language model in order to prevent the insertion of a liaison phoneme when the following word begins with a consonant.

7. Acknowledgements

The authors would like to thank the evaluation project Technolangue EVALDA-ESTER.

8. References


![Figure 4: Liaison with skip transition for word “les”.](image)