# A French Non-Native Corpus for Automatic Speech Recognition

## Tien-Ping Tan, Laurent Besacier

CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, FRANCE

Email: (Tien-Ping.Tan, Laurent.Besacier)@imag.fr

## Abstract

Automatic speech recognition (ASR) technology has achieved a level of maturity, where it is already practical to be used by novice users. However, most non-native speakers are still not comfortable with services including ASR systems, because of the accuracy on non-native speakers. This paper describes our approach in constructing a non-native corpus particularly in French for testing and adapting non-native speaker for automatic speech recognition. Finally, we also propose in this paper a method for detecting pronunciation variants and possible pronunciation mistakes by non-native speakers.

## 1. Introduction

Automatic speech recognition applications are becoming increasing popular. However, as automatic speech recognition matured, speech recognition performance on non-native speakers is still low. Non-native speakers are often given a second class treatment in term of speech recognition services. As the world become more globalize, non-native speakers are not a minority group anymore. International spoken languages for instance English and French are taught as a second language at schools and universities in most countries in the world. In addition, people who are on vacation in some foreign country also often learn up some common phrases, with the help of Internet or travel books that can be easily found nowadays.

Research in non-native speech recognition is becoming more active since the late 90s, although there are still not many compare to the works in other areas of speech recognition. A classical study on non-native speech recognition is done by (Uebler and Boros, 1999). Works in non-native speech recognition try to take into account of the way non-native speakers speak. Most of the works focus in acoustic model adaptation and dictionary improvement. In non-native acoustic model adaptation, among the pioneers are Witt and Young (1999) and Tomokiyo (2000). Adaptation method either use speaker's native language (Witt and Young, 1999), small amount of non-native language (Wang and Schultz, 2003) or the speech from the target language itself (Steidl et al., 2004). On the other hand, the work in dictionary adaptation can be found in (Goronzy, 2002, Livescu and Glass, 2000).

Speech corpora are very important components for the research and development in automatic speech recognition. It is also important to give researchers a way to compare and examine their results in a more meaningful way. However, only a few speech corpora are constructed by putting non-native speakers in mind. For instance, we can only cite non-native German corpus BAS Strange Corpus 1[1] and Verbmobil Denglish[2] which is non-native English corpus from German speakers which are publicly available. Obviously, without non-native speech samples it might be difficult to test and improve speech recognition applications.

Speech corpora generally consist of two parts-training and testing. In most cases, it is not feasible to develop a non-native corpus to train a speech recognition system, because of the difficulty in collecting enough samples and the number of possible non-native groups are simply too much. However, a more practical target will be to provide enough non-native speech samples for evaluation and research purposes. As a lot of speech corpora consist only of native speech, it is important that we integrate non-native speech with it, therefore making them more complete for non-native speech recognition evaluation and research. In this paper, we will present our methodology in constructing a non-native speech corpus for these purposes. Our work will focus on non-native French speakers particularly the Chinese from China and Vietnamese from Vietnam, although it should be applicable for any non-native group.

## 2. Speech Corpus Acquisition

The corpus is developed for testing and research in mind. For testing, we would like to test the non-native speakers in context of dialog and read articles. Concerning the domain, we have chosen the tourism domain, which might be a realistic case, where non-native speakers are likely to stumble upon.

### 2.1. Dialog

For the first part, we selected common dialog phrases in tourism domain, for example hotel, restaurant, transport and others. They were collected from web resources, travel books and elementary French language books. After the sentences were collected, we extracted the vocabularies out and used an in house pronunciation generator to generate their pronunciations. The pronunciation generation involved two steps. At the first step, we searched the pronunciation for the words with few dictionaries. Subsequently, a grapheme to phoneme application was used to generate possible pronunciation for words that were not found in our pronunciation dictionary. After the pronunciation dictionary was generated, we selected the sentences to be read by speaker for recognition from the text pool. Sentences were selected such that those with the most number of unique unseen triphones were selected, so that we can evaluate

---

[1] Available at http://www.phonetik.uni-muenchen.de/Bas/
[2] Available at http://www.elda.org/

non-native speaker in as many context as possible. The total number of unseen triphones found over time is showed in Figure 1. The graph shows that the number of unique triphones found drop dramatically for the first hundred sentences. This shows that frequent triphones are repeatedly found, which is something desirable, because they should be tested more frequently compare to rare triphones.
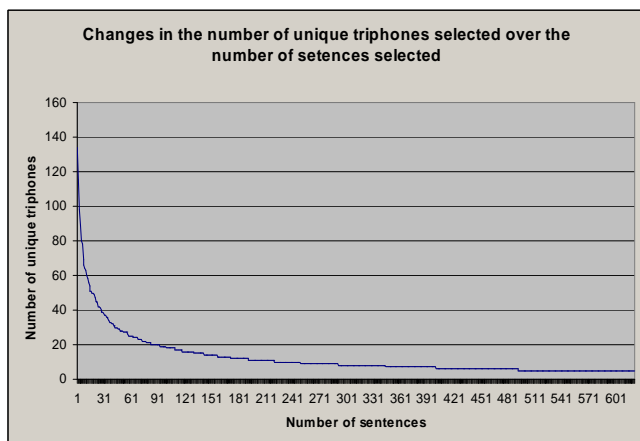


Figure 1: Changes in the number of unique triphones found over the number of sentences selected

## 2.2. Read Articles

The texts in this second part are also from tourism domain, however instead of dialog, they are sentences from tourism articles on the web. The texts were first gathered from tourism websites using a web crawler. Subsequently, we extracted the text out from the HTML files. Next, we filtered and normalized the sentences. This step involves removing punctuations, changing digits to numbers, lower case the text, changing paragraph to sentences, limiting the size of sentences etc. After manually verified that the sentences were suitable, we used the same approach described above to select sentences to be uttered by speakers.

## 2.3. Text Corpus Evaluation

To have an idea of the phone distribution in our corpus compare to the general phone distribution in French, we calculated a correlation coefficient between these two. The result shows that our corpus has a correlation coefficient of about 0.9[1] for its all three parts, which means that it is phonetically well balanced.

| | Type | Correlation Coefficient |
|---|---|---|
| 1 | Dialog | 0.910 |
| 2 | Article | 0.893 |
| 3 | Adapt | 0.920 |

Table 1: Correlation coefficients for dialog, read article and adaptation parts of our corpus

---

[1] Note here that the phone distribution only gives a general idea of the speech corpus, because we only select one possible pronunciation for each word. In addition, we assume 'liaison' occurred. This is why there is a high percentage of /z/
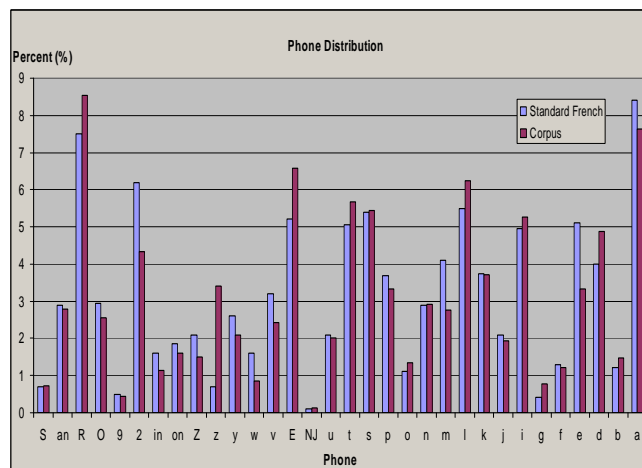


Figure 2: Phone distribution of Standard French compare with our corpus to be pronounced

## 2.4. Recording

A total of seven native Chinese speakers and eight native Vietnamese speakers with a comfortable degree of experience in the target language (French) were recruited. They are composed of seven males and eight females. Chinese speakers who took part in the recording have previously taken 500 hours of French language course in China before they came to France and they were attending French courses at the local language school, at the time of the recording. All of them have been in France for less than a year. The Vietnamese speakers are students from local universities. All of them have been in France for more than a year and have learned French for more than three years. Five of the speakers from each native group were selected to record the test part and the rest for the adaptation part. For baseline comparison, three native French speakers were also selected for recording the same test part.

Recording was done in a sound proof room, using a headset microphone, with sampling frequency of 16 kHz. EMACOP (Multimedia Environment for Acquiring and Managing Speech Corpora) was used for recording and managing the speech corpus (Vaufreydaz et al., 2000). A supervisor was assigned to monitor and facilitate the recording of each speaker.

| | French | Vietnamese | Chinese |
|---|---|---|---|
| Read Dialog | 2.84s (852s) | 3.64s (1822s) | 4.09s (2047s) |
| Read Article | 6,27s (1843s) | 10.2s (4694s) | 11.72s (5740s) |
| Adaptation | - | 12.54s (3687s) | 17.9s (3509s) |

Table 2: Average duration of a sentence and total duration (in parenthesis) of sentences read by different native groups

## 3. Speech Corpus Evaluation

Evaluating a speech corpus by analyzing every word read by non-native speakers is resource consuming. Getting phoneticians to agree upon the same transcription

is another difficulty. Since phonetic analysis of speech is not very feasible in our case, we have used an automatic time-based phoneme scoring method to give us some global information on the pronunciation behaviour of each speaker and group of native speaker as a whole. This method was allready used for creating confusion matrices for cross-lingual phone mapping (Le and Besacier, 2005).

## 3.1. Time-Based Phoneme Scoring

In time-based phoneme scoring, we measure the probability of a target phoneme mismatched with other phonemes by taking into account the time scale (see Figure 3). This will create a confusion matrix which gives the probability of a target phoneme to be mismatched with other phonemes (including itself).
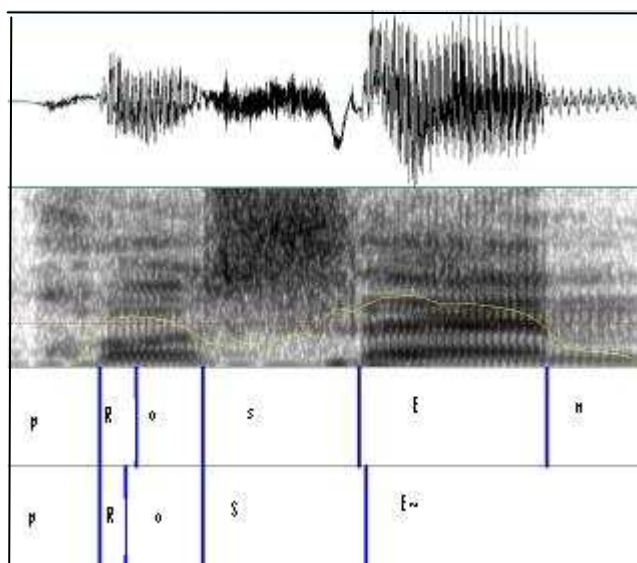


Figure 3: Time-based phoneme scoring. Tier-1 (below the spectrogram) shows the hypothesis phoneme sequence /p R o s ɛ n/ and Tier-2 (below tier-1) shows the reference phoneme sequence /p R ɔ ʃ ɛ̃/. The hypothesis shows phoneme /ɛ/ and /n/ are recognized instead of the rightful phoneme /ɛ̃/. So, the probability of phoneme /ɛ/ and phoneme /n/ replacing the phoneme /ɛ̃/ is 0.5

We performed time-based phoneme scoring by having a 'hypothesis' alignment to compare against a 'reference' alignment. In our case, we obtained the reference alignment using force alignment using only French acoustic model. But for the hypothesis alignment, phoneme recognition was done using combination of French acoustic model and the acoustic model of the speaker's native language. The multi-lingual model used was trained using BREF120 speech corpus (Lamel et al., 1991) for French phonemes, our Vietnamese corpus (Le et al., 2004) for the Vietnamese phonemes and CADCC[1] speech corpus for the Chinese phonemes. For phonemes common to different languages, a different model was built for each language (for instance we built a model for o in French and one for o in Vietnamese). It is very rare if not impossible for reference phoneme sequence and hypothesis phoneme sequence to align exactly at the same time, therefore boundaries are often overlapped. This may

---

[1] Available at http://www.d-ear.com/CCC/corpora.htm

reduce the probability and confidentiality of the result. To reduce cases like this, we set the duration threshold as 0.2. This means that the hypothesis phoneme with duration less than 20 percent of the duration of the reference phoneme will not be counted.

## 3.2. Discussion

We analyzed the two most likely hypothesis phonemes for each French reference phoneme for non-native speakers, and as for baseline comparison, the same test was also performed on native French speaker. The Table 3 below shows an excerpt of the results obtained from the phoneme alignment for a particular Vietnamese speaker.

| Hypothesis Phoneme | Reference Phoneme | Percentage |
|---|---|---|
| b (vn) | b | 0.37 |
| d (vn) | b | 0.09 |
| s | ʃ | 0.45 |
| ʂ (vn) | ʃ | 0.17 |
| z | ʒ | 0.31 |
| z (vn) | ʒ | 0.30 |
| o (vn) | o | 0.41 |
| o | o | 0.19 |

Table 3: An excerpt of the confusion matrix of a Vietnames speaker. It shows two most likely phonemes substitution for each (reference) French phoneme. Phoneme with (vn) is a Vietnamese phoneme and the one without is a French phoneme

### 3.2.1. Analysis of Native Vietnamese Speech

According to the IPA, there are 32 phonemes in French and 41 in Vietnamese. Twenty two of them are 'similar', which exist both in Vietnamese and French.

The results from the time-based phoneme scoring showed that in most cases for similar phonemes, the same French and Vietnamese variants were recognized as the two most likely phonemes for the speakers (e.g. /a/ was recognized as French /a/ and Vietnamese /a/). However, whether French variant or Vietnamese variant are stronger is speaker dependent, since some speakers have a stronger Vietnamese variant, others have a stronger French variant instead. It is also interesting to note that for phoneme /p/ which exists in French and Vietnamese (exists only as a word final unreleased stop /p/ in Vietnamese), Vietnamese variant of /t/ was recognized instead for three of the five speakers. This may indicate that the phoneme /p/ by native Vietnamese speakers is different from the one uttered by native French speakers. On the other hand, baseline results from native French speakers show only little or no Vietnamese variants in most cases.

For 'new' phonemes, which exist only in French but not in Vietnamese, a systematic substitution of phonemes by native phonemes occur in most cases, see Table 4.

There are few cases where phonemes have many different substitutions. This may indicate that for these particular phonemes, the native Vietnamese speakers have difficulty to pronounce them.

| French Phoneme | Phoneme substitution (Vietnamese) | Phoneme substitution (baseline-French) |
|---|---|---|
| ø | ʁ (vn), ø | ø |
| œ | ʁ (vn), œ | œ |
| ə | ʁ (vn), ə, ø | ə, ø |
| ã | ã, ɔ (vn) | ã, ɔ̃ |
| g | R | g |
| ɛ̃ | ɛ̃, ɛ̆(vn) | ɛ̃ |
| ʃ | ʃ, ʂ (vn), s | ʃ |
| œ̃ | œ̃, a | œ̃, ɛ̃ |
| ʒ | ʒ, z, z (vn) | ʒ, ʃ |

Table 4: Frequent phoneme substitution by native Vietnamese speakers (col 2) and native French speakers (col 3). These are the phonemes which are not found in Vietnamese IPA. The list of possible substitution appears in descending order of the frequency of substitution. Only substitutions appearing more than once are included

### 3.2.2. Analysis of Native Chinese Speech

There are nineteen similar phonemes between French and Mandarin (Duanmu, 2002). Chinese phonemes variants are also identified in many cases when similar French phonemes are expected. However in most cases, French variants seem to be stronger. There are cases of (similar) phoneme substitution by a completely different phoneme (e.g. substitution /u/ by /o/) by many speakers, this may suggest the same confusion happened. One of the possible reasons is the influence of the graphemes of the word which confuse the speakers. Like Vietnamese speakers, many 'new' phonemes are substituted by native variants. Table 5 shows part of the results from three of the six native Chinese speakers.

| Sp | YX | | YS | | LC | |
|---|---|---|---|---|---|---|
| | 1st | 2nd | 1st | 2nd | 1st | 2nd |
| Similar phonemes | | | | | | |
| a | a | a (cn) | a (cn) | in | a | a |
| i | i | e | i | i (cn) | e | i |
| j | j (cn) | j | j (cn) | i | j | tɕ (cn) |
| p | SIL | p (cn) | SIL | b | SIL | p (cn) |
| u | o | u | u | o | u | o |
| New Phonemes | | | | | | |
| œ | ʁ(cn) | in | ɔ | ə (cn) | ʁ(cn) | ɔ |
| ø | SIL | ʁ(cn) | ø | o | SIL | ʁ(cn) |
| g | k | SIL | e | k | k | kʰ (cn) |
| d | SIL | tʂʰ (cn) | d | SIL | SIL | t |
| z | z | s | SIL | z | s | s (cn) |
| R | R | kʰ (cn) | R | x (cn) | R | x (cn) |
| ɔ̃ | ã | ɔ̃ | ŋ (cn) | ɔ̃ | ɔ̃ | SIL |
| œ̃ | un | a (cn) | in | SIL | un | ʁ(cn) |
| ã | ã | ŋ | ã | ŋ (cn) | ã | SIL |

Table 5: An excerpt of the list of two most likely French phoneme substitutions by three of the native Chinese speakers. Note that, the phoneme with (cn) is a Chinese phoneme and the one without is a French phoneme

## 4. Summary

We presented in this paper an approach to build a French non-native corpus for testing and adaptation. The method can also be used for any other languages. With this corpus, in future, different methods can be tested to improve a speech recognition system. The adaptation part can also be experimented with different adaptation methods. We also propose a method for detecting pronunciation variants and pronunciation errors by non-native speakers. The method seems promising in uncovering the pronunciation pattern of non-native speakers. This initial information can be used to improve our pronunciation dictionary used in ASR or as an information for the study of second language learning (L2).

## 5. Reference

Duanmu, S. (2002) *The Phonology of Standard Chinese,* New York, Oxford University Press.

Goronzy, S. (2002) *Robust Adaptation to Non-Native Accents in Automatic Speech Recognition,* Berlin, Springer Verlag.

Lamel, L. F., Gauvain, J. L. & M., E. (1991) BREF, a Large Vocabulary Spoken Corpus for French. *Eurospeech-91.* Genoa, pp. 505-508.

Le, V.-B. & Besacier, L. (2005) First Steps in Fast Acoustic Modeling for a New Target Language: Application to Vietnamese. *ICASSP 2005.* Philadelphia, USA, pp. 821-824.

Le, V.-B., Do-Dat, T., Casteli, E., Besacier, L. & Serignat, J. F. (2004) Spoken and written language resources for Vietnamese *LREC 2004.* Lisbon, pp. 599-602.

Livescu, K. & Glass, J. (2000) Lexical Modeling of Non-Native Speech for Automatic Speech Recognition. *ICASSP-00.* Istanbul, pp. 1683-1686.

Steidl, S., Stemmer, G., Hacker, C. & Nöth, E. (2004) Adaptation in the Pronunciation Space for Non-Native Speech Recognition. *International Conference on Spoken Language Processing (ICSLP 2004).* South Korea, pp. 2901-2904.

Uebler, U. & Boros, M. (1999) Recognition of Non-native German Speech with Multilingual Recognizers. *Eurospeech-99.* Budapest, pp. 911-913.

Vaufreydaz, D., Bergamini, J., Serignat, J. F., Besacier, L. & Akbar, M. (2000) A New Methodology for Speech Corpora Definition from Internet Documents. *LREC2000, 2nd International Conference on Language Ressources and Evaluation.* Athens, Greece, pp. 423-426.

Wang, Z. & Schultz, T. (2003) Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization. *Proc. Eurospeech-03.* Geneva, Switzerland, pp. 1449-1452.

Witt, S. & Young, S. (1999) Off-Line Acoustic Modellling of Non-Native Accents. *EuroSpeech-99.* Budapest, Hungary, pp. 1367-1370.