Design of Vietnamese Speech Corpus and Current Status

Luong Chi Mai¹, Dang Ngoc Duc²

¹Institute of Information Technology, Vietnamese Academy of Science and Technology, ²Alcatel, Vietnam
lemai@ioit.ac.vn; ngoc-duc.dang@alcatel.com.vn

Abstract. This paper presents a current status and activities for spoken language resources for Vietnamese implemented in research institutions such as Institute of Information Technology, Vietnamese Academy of Science and Technology, and International Research Center MICA, Hanoi University of Technology. This is our first attempt of a process of building a large Vietnamese speech database and the corpora should be in a common design to make it available for researchers in Vietnamese speech processing.

Keywords: Speech corpus, Design of speech corpus, General corpus, Specific corpus, Phonetic structure, Data collection, Recording, Labeling.

1 Introduction

Research in Vietnamese speech synthesis and speech recognition began nearly two decades ago. However, very few studies have been carried out on the Vietnamese speech research. One of the obstacles on this sector is that there is not standard and reliable Vietnamese corpus. Corpus is used to all processes of speech recognition system: train the basic recognition units, build the language model, do force-align utterances, embedded training, and recognition testing... The quality of corpus is important to make the speech recognition research to navigate in the good direction and then to obtain the good achievements. However, building a corpus is always the time-consuming and big-investment task. In Vietnam, each research entity makes corpora itself and due to limitations of budget, time and knowledge these corpora are always small and not high quality. In addition, isolated corpora make co-operation among research units very difficult or impossible. Building a complete, high-quality Vietnamese corpus for common use in all research units in Vietnam becomes an urgent demand to eliminate obstructions in availability of Vietnamese corpus and to boost the research in this field. In the Session 2 some basic structure of Vietnamese language is presented. Session 3 of the paper describes a solution, proposed to build a corpus which has multi-speakers, full-context phonemes sentences with high-quality recording and phonetic labeled. This corpus can be used for research general continuous independent-speaker recognizers as well as specific-purpose recognizers such as digit recognizer, name recognizer, ... At the end of the session will describe some corpus that follows the design. Collection and labeling are processed at the
Institute of Information Technology, Vietnamese Academy of Science and Technology. Session 4 presents strategy, research for building large corpus and data collection at the International Research Center MICA, Hanoi University of Technology. By supporting of the National Project on Vietnamese Language and Speech Processing 2006-2008, we are trying to collaborate with each other to make a common one for the speech community. Session 5 is a conclusion.

2 Basic Phonetic Structure of Vietnamese

Vietnamese language is a mono-syllabic and tonal language with 6 tones. A syllable in full structure (a tonal syllable or an isolated word) has five parts: initial sound (consonant), medial sound (semi-vowel), nucleus sound (vowel or diphthong), final sound (consonant or semi-vowel) and tone (see figure 1). Except the initial consonant (called INITIAL part), the rest of the syllable is called a FINAL part. Tonal syllable (6,492) Base syllable (2,376) FINAL (155) INITIAL (22) Medial (1) Nucleus (16) Ending (8) Tone (6). Vietnamese is a monosyllabic language (each syllable is one isolated word). There are about 19,000 pronounceable distinct syllables in Vietnamese [1]; however there are only about 6,700 isolated words. From this isolated-words vocabulary, we have firstly extracted all 22 initial parts, 155 final parts and 6 tones [1]. The general structure of a Vietnamese syllable is described in the Figure 1.

<table>
<thead>
<tr>
<th>Tone</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Onset</td>
</tr>
<tr>
<td>Initial</td>
<td></td>
</tr>
</tbody>
</table>

Figure1. General structure of a Vietnamese syllable

The tone is super-segment and has on all parts of a syllable. There are six distinct tones in Vietnamese and they can affect the meaning of individual words. The following Figure 2 describes six Vietnamese tones denoted from 1 to 6.

Figure 2. Six tones in Vietnamese
Syllables with a closure coda can only go with rising tones and drop tones. As 7th and 8th contour patterns in Figure 2, rising and drop tones of syllables ending with stop consonants have F0 contours similar to rising and falling tones of other syllables, but they rise or drop more sharply. Therefore, most linguists who study Vietnamese acoustics claim that the Vietnamese language contains 8 different tones base on F0 contours.

Initial is always consonant; it may be omitted in some syllables. There are 21 initials in Vietnamese:

- Seven plosives: b, d, t, ð, c, k;
- Four nasals: m, n, ŋ, ŋ;
- Nine fricatives: f, v, s, ñ, ñ, ñ, h; one lateral: l.

There are 155 finals in Vietnamese and the final may be decomposed into onset, nucleus and coda. Onset and coda may be optional; they may not exist in a syllable. Nucleus is a vowel or a diphthong; coda is a consonant or a semi-vowel. There are two onsets, 16 nucleus and 8 codas in Vietnamese as listed hereby:

- Onset: ŋ, zero
- Nucleus: i, e, ê, u, o, ō, ù, y, ň, a, å, ie, uo, wo, w
- Codas: p, t, m, n, k, ŋ, ŋ, ŋ

Initial, tone, onset, nucleus and codas may be combined together to make a syllable. However, not all combinations are possible, for example the onset /u/ is always not in front of the lip-rounded vowel: /u, o, ō, ŋ/.

3 Design of Vietnamese Speech Corpus

The corpus is being designed and developed at the Institute of Information Technology, Vietnamese Academy of Science and Technology and being supported by National Project on Vietnamese Language and Speech Processing of the period 2006-2008. However for the time being, only small part of the corpus is done and it will be described in the last sub-session of this part.

The main goal is to the design and realize of a database to be used in an automatic speech recognition system as well as speech synthesis. The objective is two-fold. On the one hand, the database is to provide the Vietnamese researchers with a basic amount of speech material for general speech research, including speech synthesis and speech recognition. On the other hand, the speech corpus is designed to obtain a controlled amount of speech for developing commercial speech recognition engines in given purposes such as number recognizer, limited command recognizer, name recognizer, ...

The corpus consists of two parts. The first one contains data for training conventional context-dependent phone models and data for testing, which is independent-speaker speech with training data. This part aims to be used for general research on algorithms for phonetics dictionary, phonetics labeling, training model, classifying, language model, noise cancellation...Two types of utterance are defined in this part: the connected words and isolated word. The isolated words are useful to study on isolated word recognition system and to train the mono-phone models. The sentences with connected word will be used in continuous recognition systems.
The second part is sets of sentences with different specific purposes. For example, set of sentence of continuous digit, set of sentence of names of people... The recognition system trained with data in the first part will be customized with each set in this part to develop the commercial application on speech recognition.

Spoken language corpora as research tools should be designed and collected defining and bearing in mind what their use would be. The primary purpose is the development of speaker independent continuous speech recognition in Vietnamese that means a large speaker population. Another desirable purpose is the development of specific applications, such as the digit continuous recognizer, proper name recognizer.

It is the first Vietnamese database of its kind, and there is no previous (to our knowledge) experience with modeling units for continuous speech recognition in Vietnamese, the database should be used for acoustic modeling at the phoneme level, and together with an as uniform as possible coverage of age and sex groups. This criterion is used to design the text prompts read by the speakers, which contains a controlled number of occurrences for each of a predefined set of acoustic modeling units so that reliable model parameters could be estimated. The database contains two parts:

**Part 1. Design of General Corpus.** This corpus will be used for general purpose to research continuous independent-speaker recognition with the large vocabulary. The number of speakers is about 200-300 people, in which 50% are male and 50% are female. The ages of speakers are distributed from 15-45. The sentence is designed so that the context balance is obtained among Vietnamese phonemes. The number of sentences is about 300 sentences, each sentence is spoken by one speaker for at least 3 times. The size of vocabulary is about 2000-3000 syllables in Vietnamese.

**Part 2. Design of Specific Corpus.** This corpus is used to make the recognition applications Continuous digit corpus. Used for building continuous digit recognizer. The number of speaker is between 100-200 people, in which a half is male and the other is female. The ages of speakers are from 15-45. The concurrence of digits should be approximately the same for all digits. The number of word in corpus will be from 10000 words. The sentences in corpus consist of digits with random order and have variant lengths. Each sentence is recorded at least 3 times for each speaker. Name corpus consists of popular Vietnamese names including family name and first name. The number of speaker is between 100-200 people, in which a half is male and the other is female. The ages of speakers are from 15-45. Each sentence consists of one full name and spoken by a speaker for at least 3 times. The sentence should contain as many as possible names of people in Vietnam. The size of vocabulary is estimated about 2000 words.

### 3.1 Labeling

Labeling is process defining the boundaries of acoustics unit in an utterance. There are some levels of labeling: sentence, syllable, and phoneme. The labeling can be done in two modes: by hand or automatically by computer with force-align process.
The hand-labeling is recognized as more accurate than the automatic labeling. However, this method has two disadvantages: it takes about 11-30s for labeling each phoneme; hence labeling a corpus requires a huge work and very long working days. The other disadvantage is there are disagreements between labeler, some papers have pointed that the difference among professional labelers is about 30% in term of narrow phonemes.

The automatic labeling is much faster than hand-labeling, but it does not have the same accuracy as the first one. However, this method can be used for labeling vast data of a corpus in short time.

**Machine-readable Transcription System**

Transcription is used for encode the signal file into the text file in some level: syllable level and phoneme level. For the time being, there is not a standard ASCII encoding transcription of Vietnamese phonemes. The SAMPA system is used widely in the world as standard for transcribing. Unfortunately, the SAMPA for Vietnamese has not been developed yet. The IPA conversion has been used throughout the world as a standard for phonetic transcription and may be used in Vietnamese. The disadvantage of IPA is that IPA contains non-ASCII characters, which make the coding difficult. The Worldbet [3] is proposed as a standard for phonetic transcription for multilanguage using ASCII characters. However, Worldbet label set for Vietnamese contains some differences with traditional phonetic Vietnamese pronunciation. To solve this problem, a label set with ASCII characters for Vietnamese phonetic transcription is developed based on the IPA and Worldbet idea. This system is based on the International Phonetic Alphabet (IPA) to include notations at segmental tonal levels which are phonemic, plus notations at the prosodic levels. The Telex typing manner is proposed to be used for both syllable transcription as well as phonemes transcription. The table in Annex 1 shows the ASCII encoding and corresponding phonemes.

In addition to the phoneme transcription, extra marks to aid in the interpretation of speech and non-speech events in the utterances need to be also defined. Non-speech acoustic events are arranged into four categories:

- filled pause (e.g. uh, mm)
- speaker noise (e.g. cough, grunt)
- background noise, i.e. noise that is not intermittent, and has a more or less stable amplitude spectrum
- intermittent noise (e.g. baby crying, phone ringing).

For noise events that occurred over a span of one or more words, the transcription indicated the beginning of the noise, just before the first word.

Besides the transcription encoding, an orthographic lexicon needs to be generated. The lexicon file is an alphabetically ordered list of distinct lexical items (essentially words) with the corresponding phoneme transcription.

**Labeling Strategy**

The method chosen to label these corpora is semiautomatic manner, including three steps: manual transcription of the signal files, automatic alignment of the phoneme
labels in the transcriptions with the signals, and manual verification and correction of the aligned labels. The manual transcription is process, in which the labeler hears utterances and makes corresponding transcription in syllable level. The transcription must contain only ASCII characters.

Automatic alignment is carried on by a recognizer to force-align the utterances and make time-label for each phoneme in the sentences. The force-align is similar to the recognition process, in which the inputs are signal files, the corresponding transcription. An additional process is done before force-align is transform the syllable transcription into the serial of phonemes.

Manual verification and correction. After the automatic alignment, the time-labels of phonemes are created, but due to the limitations of force-align process these labels do not have high accuracy. These labels need to be verified and corrected by labelers manually. This process is similar to hand-labeling but it needs much less time because labelers need to do only a final correction by hand of label identities and segment boundaries.

### 3.2 Corpus Implementation

#### 3.1.1 Recording

The recording platform will be based on a Pentium PC computer equipped with a Dialogic board. It is necessary that the recording software include a voice/silence detector. For each item to be recorded, the minimum initial silence, maximum initial silence, final silence and maximum recording time are specified before. If any of these conditions are not satisfied, the system requested a repetition of the item. The microphone will be placed about 25 cm from speaker’s mouth, 30 degrees off axis. Speech files will be stored as sequences of 16-bits 16 kHz A-law uncompressed speech samples (CCITT G.711 recommendation). Each prompted utterance is stored in a separate file.

In order to preserve consistent quality along all the recording period, the signal files collected in a recording session will be checked immediately for a number of quality parameters: DC bias, signal clipping, signal and noise levels, signal-to-noise ratio, and main related noise components.

The following are steps to develop the database:
- Linguistic definitions: text corpus, selection of dialect areas, phonetic transcription
- Technical setup: recording equipment, backup, formatting
- Recruitment of a large number of speakers. In order to ensure that database size goals will be met, 35% more than the target number of speakers will be contacted. Speakers will be chosen according to the dialectal distribution mentioned above.
- Recording data
- Orthographic transcription and labeling of the recorded speech.
3.1.2 Corpus Organization

The database consists of several sub-databases:

1. /general \{Material for training, testing the general recognizer\}
2. /Digit \{Material for training, testing digit continuous recognizer\}
3. /Name \{Material for training, testing proper name recognizer\}
   ... \{Other sub-database to be inserted depending on the specific purposes\}

Each sub-database has directory hierarchy as follows:

1. / \{database root directory\}
2. /train \{to be used for system training\}
3. /test \{material to be used for system testing\}
4. /doc \{online documentation and tables unusual\}

The “train” and “test” directories contain sub-directories corresponding to each speaker, whose names are coded as follows: XXXSRR where:

XXX : speaker identifier
S : sex code: F for female, M for male
RR : region code. For example: HN for Hanoi, SG for Saigon, HU for Hue,

Each sentence directory contains 3 sentence-related files.

The file types are as follows:

wav – wave file
phn - phonetically-based transcription
txt - orthographic transcription (prompt form)

3.1.3 Automatic labeling

A first step towards this is to make a recognizer. The three-state left-to-right sex-dependent HMMs with Gaussian density mixture output probability functions is proposed to be used. To obtain a more rapid convergence of the training process, the phonemically compact sentences labeled manually are used for initialization of mixture parameters followed by a Baum-Welch re-estimation of both mixtures and transition probabilities, and a concatenated training used all available signals and associated transcription files. To obtain a good time resolution, the acoustic processing used frames 10 ms long, from which 26-dimensional vectors including 12 PLP coefficients, log energy, and their first derivatives, are computed.

Automatically aligned label files are obtained through a Viterbi decoding of each signal file guided by a network generated from the associated transcription.

3.1.4 Hand-labeling

The labeler will be organized into at least two groups. Each group does the labeling independently. They verify and correct labels using a simultaneous synchronized display of signal waveform, spectrogram, and labels, based on listening and visual examination. To ensure consistency, a display at a constant 1 ms/pixel resolution is used, and rules are defined to be used in cases where different boundary placements are arguable. After finished, two groups will do cross-check to diminish the disagreements between groups.
3.3 Data Collection

3.3.1 Telephone Number Corpus

a. Description
   i. Vietnamese digits are monosyllabic words read differently depending on the context, i.e. the digit position in the sequence. For example, digit 4 is read as "boons" at the beginning and possibly read as "tuw" at the end of the sequence.
   ii. Telephone numbers (hereinafter called numbers) are sequences of nine to ten digits.

b. Statistics
   i. Method of recording:
      - For Mobile phone: Mobile phone microphones in outdoor environment with average signal to noise ratio (SNR). Each speaker read a number three times. Audio format is WAV with 8 kHz sampling rate and 16-bit resolution. Size of corpus: 1541 numbers.
      - For stable telephone: The corpus was digitized at an 8000 Hz sampling rate with A/D conversion precision of 8 bits. Speakers recited their telephone numbers, street addresses, ZIP codes or other numeric information over the telephone network in a natural speaking manner. The data were collected from different environments and may contain a noticeable amount of noise and other "real-life" aspects such as breath, glottalization, and music.
   ii. Number of speakers:
      - For Mobile phone: 170 including 94 males (app. 55%) and 76 females (app. 45%) from various localities in the North of Vietnam. Hence, the corpus consists of mostly non-standard Vietnamese accents. The standard is assumed to be Hanoi accent.
      - For stable telephone: 208 including 130 males and 78 female from mostly from the south of Vietnam. Corpus consists of 442 sentences with 2340 words [2]

c. Labeling: The corpus, at syllable level, was labeled manually at phonetic level, using forced alignment with manual adjustment afterward.

3.3.2 Broadcasting Speech Corpus VOV (Voice of Vietnam)

a. Description
   The corpus, including stories, mailbags, reports, etc. broadcast on the Voice of Vietnam [5], was collected from about 15 speakers (one male and one female) with standard Vietnamese accent. Sound files in RealAudio format were collected from VOV website and converted into WAV format (bit rate 256 kbps, mono channel, sampling rate 16 kHz).

b. Statistics
   The corpus contains 29062 utterances of on average 10-syllable length. The number of distinct syllables with tone is 4379; the number of distinct syllables without tone is 1646 covering most of Vietnamese syllables. The total capacity of the corpus in WAV format is about 2.5GB.
One deficiency of the corpus is the number of speakers. In one radio station, there are only a limited number of broadcasters. Their voices do not cover most variations of Vietnamese speech. The corpus is also not phonetically balanced. The data gathered from the section of story reading is the largest part of the corpus with about 1 GB.

c. Labeling:
Long sound files were chunked into utterances using a silence detector. The corpus was listened and manually labeled at syllable level.

3.3.3 TTS Corpus

a. Description
The corpus content was extracted from a short story. Reader was a female with standard Vietnamese accent. Recordings are of WAV format, 11 kHz sampling rate, and 16-bit resolution.
b. Statistics
The corpus contains 567 utterances of an average length of 15 syllables.
c. Labeling
Syllable level transcription of the whole corpus was completed. Phoneme level transcription with tone information, used for synthesis experiments, was done for 20 utterances.

4 MICA VNSpeechCorpus

4.1 Text Corpus
Two phases of collecting text data were implemented in MICA project [3, 4]. In the first phase, data is collected by some experts in order to ensure the desired requirements. And in the second phase, data is extracted automatically with one desired distribution of acoustic units from the web corpora as a follow: First, by using a web-robot (or web-spider), web pages are collected and stored and then, these web pages were filtered and analyzed for building a text corpus. Finally, a language model was estimated from this text. There are about 2500 Vietnamese websites in Vietnam which publish: daily news, information, entertainment, e-commerce, forum... The daily news web pages introduced a constraint in the data collection, since we had to regularly access the same sites to get an acceptable amount of data. This is the major difference with web data collection for a majority language like French or English where there are enough web pages that can be collected at a given time.

Beside database of phonemes, digits, application words, other data including sentences and paragraphs were collected from different resources such as stories, books, and web documents... The selected data covers different fields and contains many dialogs and short paragraphs. This initial data then was manipulated and was divided into smaller paragraphs and conversations (about 4-6 lines/paragraph or conversation) that help speaker to utter or read easily.
4.2 Corpus Organization

The VNSpeechCorpus contains 5 different kinds of data: Phoneme, Tones, Digits and string of digits, Application words, Sentences and paragraphs.

The phonemes are read by all speakers. The vowels can be read independently except two vowels á and â because their sounds are only represented completely in words in which they appear, such as-ngán (short), tân (new). The consonants are combined with vowel ơ and falling tone for pronouncing.

Vietnamese is a tonal language with 6 tones. The speakers are asked to read the words with different tones. These words have almost the same initial and final part, but they have different tones.

The digit corpus consists of isolated digits, connected digits and natural numbers. In Vietnamese digital system, most of the digits and numbers are read or uttered with the unique sound. However, there are some synonyms, especially, the numbers ended by digit 4 and 5; they could be read in different ways. In order to cover all cases, the corpus consists of all of the variants (synonyms) of theses digits and numbers.

A set of more than 50 application words is defined in the corpus. Each word corresponds to an action which is useful in several applications such as telephone services, measurement, human-machine interface ...

After selecting and processing selected paragraphs and conversations, the sentences corpus is divided into two parts, a common part and a private part. The common part contains 33 conversations and 37 paragraphs. They were read by all speakers. The private part includes about 2,000 short paragraphs, each speaker was asked to read 40 paragraphs.

4.3 Distribution of Acoustic Units

To evaluate the corpus, several modules are used to analyze the distributions of acoustic units including mono-words, base syllables, Initial-Final parts, phonemes and tones in the corpus and compare their distribution with the distribution obtained on a larger corpus which was collected in Web corpus. The acoustic units distribution obtained on this large web corpus is considered as the reference distribution of what can occur on Vietnamese language. Vietnamese is a monosyllabic and tonal language. Besides analyzing the distributions of phonemes (mono-phone, diphone and triphone) like other language as it have shown in Figure 3 [3, 4], an analysis of the distributions of tones, initial and final parts is carried out.

![Figure 3. Distribution of mono-phones in common part, private part and Web corpora](image)

In addition, the coefficients correlation is calculated between the distributions of the common part and the private part with the web reference corpora. In [3, 4] shows
that the correlation coefficients are near to 1 and it implied that the corpus is acceptable and correctly balanced in terms of acoustic units and tones.

4.4 Speaker selection and recording

50 speakers (25 females and 25 males with the age from 15 to 45 years old) are most educated at university level and represent 3 major dialect regions: the South, the North, and the Central. Each speaker has been asked for recording about 60 minutes, which includes 45 common minutes of phonemes, tones, digits and strings of digits, application words and common sentences and paragraphs corpus, and 15 private minutes of about 40 short paragraphs. For the acquisition, and managing of speech signals during recording, the EMACOP system is used and developed at CLIPS. EMACOP is a Multimedia Environment for Acquiring and Managing Speech Corpora, running under Windows 9x and Windows NT. EMACOP meets SAM specifications on input and output. In MICA project, recordings will take place with a SENNHEISER HMD 410-6 head microphone and a microphone pre-amplifier Soundcraft Spirit Folio FX8. The sampling frequency is 16 kHz.

5 Conclusion

In this paper we presented our attempt to building speech corpora and the design that we want to collaborate in the framework of National Project on Vietnamese Language and Speech Processing from the period 2006-2008 to make the common corpus for speech community doing research on Vietnamese speech recognition and synthesis. The individual corpora are being developed at the Institute of Information Technology, Vietnamese Academy of Science and Technology and International Research Center MICA, Hanoi University of Technology.

References