simultaneously satisfy these requirements. Therefore, building of Vietnamese corpus VDSPEC (Vietnamese Dialect Speech Corpus) was studied to meet the requirements for speech recognition and Vietnamese dialect recognition. It is known that dialect is a form of the language spoken in different regions of the country. These dialects may have distinctions of words, grammar and pronunciation modalities. For Vietnamese, researches on dialects are mainly concentrated on language approach [4]. In our research, we focus only on pronunciation modality for voices of Hanoi and Hue and the dialect identification is based on signal processing, hence the corpus does not reflect the difference of dialect words and grammar between these regions. Vietnamese is a tonal language. On the other hand, the tones of Vietnamese play a very important role in Vietnamese because they take part in the meaning of the word. The pronunciation modality of Vietnamese tones differs for different dialects. Therefore, the analysis of this pronunciation modality has an important implication in the identification and synthesis of Vietnamese dialects.

Section 2 of this paper will present the methods for building Vietnamese corpus in which different topics are recorded to take account of tonal balance for some Vietnamese dialects. Section 3 describes in detail the corpus and the statistical analysis of F0 variation of dialects in this corpus. Finally, section 4 gives conclusions and development in future.

2 Method for building Vietnamese corpus

There are already dialectal corpora for some languages such as English [5], Chinese [6], Arabic [9], Thai [11]... For English, FRED is really a big dialect corpus which cover 8 dialects with 2.45 million words of text and about 300 hours of speech. FRED contains data from 420 different speakers, the age of speakers included in FRED ranges from six years to 102 years. For material included in FRED, it was recorded over 30 years. The corpus permit the investigation of phenomena of non-standard morphosyntax beside analysis of phonetic or phonological details. For Chinese, there are eight major dialectal regions. The authors in [6] have built the corpus for Wu dialect belonging to eight major Chinese dialects and providing information at four levels: phonetic level, lexicon level, language level and acoustic decoder level.

Our corpus is built mainly for the first step research on dialect identification of Vietnamese and the corpus's target is more modest and meets the basic criteria. The corpus is built to cover a relative large range of topics, text contents ensure tonal balance, gender equilibrium for speakers, speakers are selected so that they possess local accent and their voices are steady, low noise for recording environment. For a corpus, there are two ways for recording: spontaneous speech and read speech. To be more active, we have chosen read speech for recording.

The building of Vietnamese corpus is done in two stages. Stage 1 includes compilation, collection and classification of documents by topic; performing adjustments to ensure tone balance in the prepared text. Next, in stage 2, recording is performed using specialized equipment with selected environment. The following is description in detail for these stages.

The topics are selected from electronic documents. The words of these topics need to be counted to ensure tone balance. Tone balance means that the appearance probability of six tones is the same in quantity (about 717 words for each tone). This procedure is conducted automatically with the support of software or manually.

The topics include life sciences, business, law, cars, motorcycles, texts are collected from electronic media VnExpress. 150 sentences containing 4333 syllables have been collected, classified and selected.

The selection of speakers have a significant impact on the quality of obtained voice. Speakers are chosen so that they speak with the local accent. The average age of speakers is 21 year old. At this age, voice quality is steady with full features for local voice. The recording is also held in different sections to cover the voice variability of human being.

Audio is recorded as standard PCM, uncompressed, with sampling frequency of 16 KHz, 16 bits per sample with one channel (mono).

3 Results

The corpus consists of 50 male voices and the same for female voices. There are two main dialects of Vietnamese for the corpus. The number of northern dialect speaker is 50 and the same speaker number for middle dialect. For each dialect, the number of male voices is equal to the number of female voices. In our case, northern dialect is Hanoi voice and middle dialect is Hue voice. For a topic, each speaker reads 25 sentences in total. The number of recorded sentences is 15000 (100 speakers and 150 sentences for a speaker). The corpus capacity is 3.62GB and total duration is 33.79 hours.

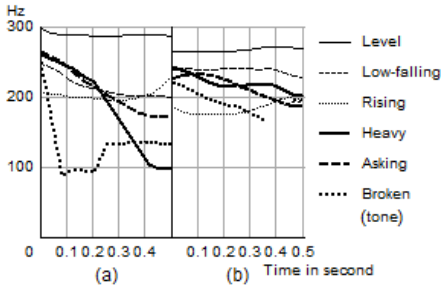

Fig. 1. Variation of 6 tones for female voices. (a) Hanoi, (b) Hue

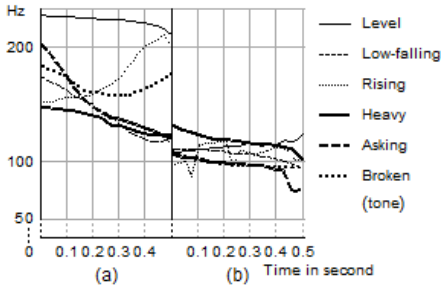

Fig. 2. Variation of 6 tones for male voices. (a) Hanoi, (b) Hue

Praat [8] was used to estimate fundamental frequency variations for Vietnamese tones in VDSPEC and four representative voices including 2 males and 2 females with two dialects were selected. The durations of the actual tones are usually different. To make the difference more evident, these durations have been normalized by the same interval 0.5 seconds. The results are shown in figures 1 and 2.

For level tone, F0 variation is rather small at around the mid level for both dialects. For Hanoi voice, rising tone starts as mid and then rises but for Hue voice the difference between starting and ending values for F0 is smaller than Hanoi voice.

For low-falling tone, F0 starts low-mid and falls monotonously. With heavy tone, F0 starts mid or low-mid and rapidly falls at the end for Hanoi voice. For asking tone (falling rising tone), F0 goes down and has a tendency to goes up at the end with Hanoi voice. With broken tone, F0 falls down, maybe is broken before going up for Hanoi voice. In general, F0 of tones for Hue voices has tendency to go down monotonously as low-falling or heavy tones for Hanoi voices.

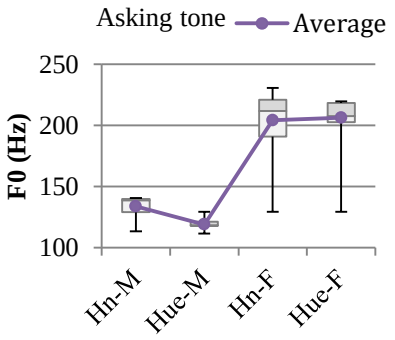

Fig. 3. F0 variation for asking tone

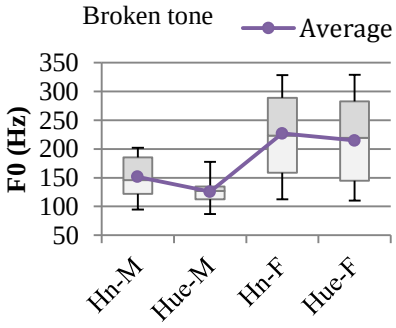

Fig. 4. F0 variation for broken tone

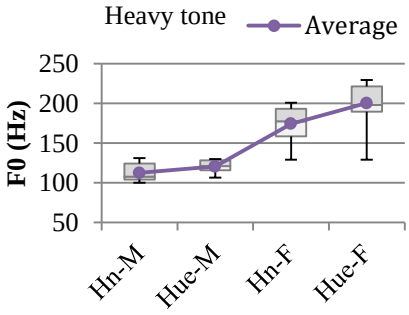

Fig. 5. F0 variation for heavy tone

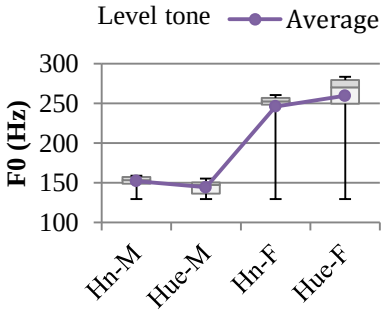

Fig. 6. F0 variation for level tone

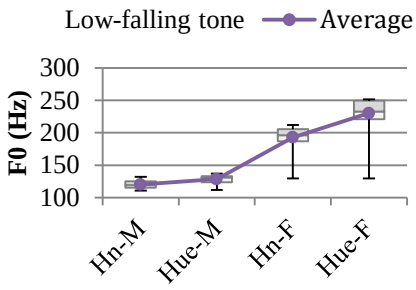

Fig. 7. F0 variation for low-falling tone

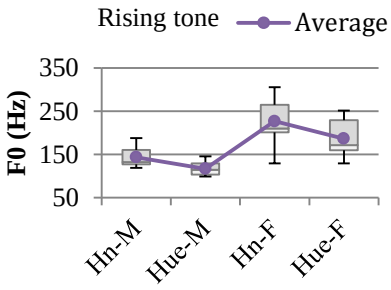

Fig. 8. F0 variation for rising tone

The variation of F0 values for 100 speakers including 50 males and 50 females is also evaluated and is depicted by boxplots in Figures from 3 to 8. These figures show F0 variation for Hanoi male voices (Hn-M), Hanoi female voices (Hn-F), Hue male voices (Hue-M) and Hue female voices (Hue-F). For each dialect, the number of female voices equals 25 and the same for the number of male voices. From Figure 3, the range of F0 variation for asking tone of Hue voices is smaller than the case of Hanoi voices, nevertheless this range for level tone of Hue voices is larger than Hanoi voices (Figure 6). For broken and rising tones, F0 of Hue voices tends to go down lower in comparison with Hanoi voices as in Figures 4 and 8. In contrast, for heavy and low-falling tones, F0 of Hue voices tends to go up higher than Hanoi voices as we can see from Figures 5 and 7. Generally speaking, the direction and the range of F0 variation for Hue tones tends to be opposed to Hanoi tones. This conclusion is also consistent with the perception in reality of the difference between the pronunciation modality for the tones of Hue voice in comparison with Hanoi voice.

To determine the signal-to-noise ratio of VDSPEC, the influence of background noise on speech signal is assumed to have properties of addition noise. This assumption is consistent with the actual condition in the recording studio. Therefore, the determination of signal-to-noise ratio is the following. During silence, which means no voice and there is only background noise, the noise power will be calculated according to the following formula:

$$P_N = \frac{1}{N}\sum_{n=0}^{N-1} b^2(n) \quad (1)$$

where P_N is short time power for the background noise, N is window length, $b(n)$ is background noise. With the sampling frequency 16000 Hz, N is selected by 256. Being based on assumptions of addition noise, the spectrum subtraction method has been implemented and we get the clean speech signal. The power of clean speech signal is calculated as follows:

$$P_S = \frac{1}{N}\sum_{n=0}^{N-1} x^2(n) \quad (2)$$

Where P_S is short time power of clean speech signal $x(n)$. Finally. the signal-to-noise ratio in dB will be:

$$SNR_{dB} = 10log_{10}\frac{P_S}{P_N} \quad (3)$$

According to the mentioned method, the signal-to-noise ratio of the corpus VDSPEC was determined and the average value of this ratio is approximately 35 dB. This value is perfectly appropriate for dialect identification and speech recognition systems.

4 Conclusions and development

This paper presents the methods and results of building a new corpus for Vietnamese taking account of tonal balance for speech recognition and Vietnamese dialect identification. The statistical analysis for the variation of fundamental frequency shows that there are distinctions in pronunciation modality of tones for Hue and Hanoi voices. These distinctions can be used as the important features in combination with other features for identifying the dialects. Our corpus will be served not only for research on dialect identification but also for Vietnamese synthesis. This corpus can be developed more completely by adding different voices and other Vietnamese dialects in the near future.

References

1. V.B. Le, D.D. Tran, E. Castelli, L. Besacier, and J-F. Serignat: Spoken and Written Language Resources for Vietnamese. In LREC 2004, Lisbon, Portugal, May 26-28, (2004), vol. II, pp. 599–602
2. T.T. Vu, D.T. Nguyen, M.C. Luong, and J-P. Hosom: Vietnamese Large Vocabulary Continuous Speech Recognition. In INTERSPEECH (2005), Lisbon, Portugal, September, 2005.
3. Vu, Q., Demuynck, K., Compernolle, D.V: Vietnamese Automatic Speech Recognition: the FlaVoR Approach. ISCSLP 2006, Kent Ridge, Singapore (2006).
4. Hoàng Thị Châu: Phương ngữ học tiếng Việt. NXB Đại học Quốc gia Hà Nội (2009).
5. Bernd Kortmann: A Comparative Grammar of British English Dialects. Walter de Gruyter (2005)
6. Jing Li et al.: A Dialectal Chinese Speech Recognition Framework. Journal of Compute. Sci. & Technol., Vol. 21, No. 1, pp. 106-115, Jan (2006)
7. Theatre Supplies and Services, http://adena.co.nz/theatre/products/sound/microphones-wired/shure/sm-series/shure-sm48.htm
8. www.praat.org
9. Fadi Biadsy, Julia Hirschberg: Using Prosody and Phonotactics in Arabic Dialect Identification. Interspeech, Vol. 1, pp 208-211 (2009)
10. Jean-Luc Rouas: Automatic prosodic variations modelling for language and dialect discrimination. IEEE Transactions on Audio, Speech and Language Processing, V. 15, N. 6, p. 1904-1911 (2007)
11. Sittichok Aunkaew, Montri Karnjanadecha, Chai Wutiwiwatchai: Development of a Corpus for Southern Thai Dialect Speech Recognition: Design and Text Preparation. The 10th International Symposium on Natural Language Processing, October 28-30, (2013), Phuket, Thailand

