International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-8,August 2016
ISSN: 2395-3470
www.ijseas.com

Statistical Analysis of Vietnamese Dialect Corpus and Dialect Identification Experiments

Pham Ngoc Hung[1,2], Trinh Van Loan[1,2], Nguyen Hong Quang[2]
[1] Faculty of Information Technology, Hung Yen University of Technology and Education, Hungyen, Vietnam
[2] School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam

Abstract
The performance of speech recognition systems will be improved if the corpus is organized in the specialized domain and is applied in a consistent way for speech recognition in specific situations. Vietnamese dialects are various. The building of corpus for Vietnamese dialect is the first step for implementing the system of dialect identification used for increasing the performance of Vietnamese recognition in general. This paper presents a method of building a corpus for Vietnamese dialect identification. Vietnamese corpus VDSPEC is built with topic-based recording and tonal balance. The duration of the corpus is 45.12 hours in total. The basic characteristics and preliminary evaluations of the corpus are also described. The statistical analysis of F0 variation and experiments on the classification of dialects using LDA projection showed that there are distinctions of pronunciation modality of Vietnamese for three dialects Hanoi, Hue and Ho Chi Minh city. For experiments on Vietnamese dialect identification, the first four formants, their bandwidths, and F0 variants have been used as input parameters for GMM. The experiment results for the dialect corpus of Vietnamese shows that the recognition rate is 66.3% without F0 information and this recognition rate increases to 72.2% with F0 information.
Keywords: *Vietnamese, corpus, Vietnamese dialects, statistical analysis, fundamental frequency, topic-based recording, tone balance, LDA projection, MFCC, formant, bandwidth, GMM, identification.*

1. Introduction

Vietnamese is a tonal language with many different dialects. It is the diversity of Vietnamese dialects that remains a great challenge to the systems of Vietnamese recognition. In other words, the pronunciation modality of the word is not the same from locality to locality. For example, for two Vietnamese dialects, the sound may be heard as the same but the sense is quite different depending on the dialect. This can reduce the performance of recognition systems if these systems have no information and training data of each dialect to be recognized.

For Vietnamese, the studies on dialects have been carried out for a long time ago but mainly on the linguistic approach and were still limited to the signal processing approach. Therefore, the research and the solution for Vietnamese dialect identification are quite necessary to improve the performance of Vietnamese recognition systems.

To be able to carry out research on speech recognition in general and in particular on dialect identification, we need a good quality corpus which meets research requirements [1], [2], [3]. For Vietnamese, some corpora exist already such as VNSPEECHCORPUS [4], VOV (Voice of Vietnamese) Corpus [5] or VNBN (United Broadcast News corpus) [6]. In dialect recognition, especially for Vietnamese language, corpus should involve the characteristics of the Vietnamese language. The mentioned available corpora do not simultaneously satisfy these requirements. Therefore, the building of Vietnamese corpus VDSPEC (Vietnamese Dialect Speech Corpus) was studied to meet the requirements for speech and Vietnamese dialect recognition.

The construction of corpus can be done in several different ways. For example, using the available audio sources from radio, television, and then classify, extract audio signals matching requirements, browse and edit the text, respectively [5], [6]. The alternative is to perform recording environments and to select speakers based on record scenario prepared in advance.

Section 2 of the paper describes the overview of Vietnamese dialects. Section 3 will present the methods for building Vietnamese corpus. Section 4 describes in detail the statistical analysis of F0 variation. Preliminary classification of dialects using LDA will present in section 5. Section 6 gives the

research results on Vietnamese dialectal identification based on GMM (Gaussian Mixture Model) using formants, their bandwidths and tonal features through the variation of fundamental frequency. Section 7 is the effect of Gaussian component number on dialect recognition performance. Finally, section 8 is conclusions and development orientations.

2. An overview of the vietnamese dialects

It is known that a 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. Vietnamese is the language that has many dialects.

The division of the Vietnamese dialects has been done by Vietnamese linguists with some different opinions. Nevertheless, the majority of linguists think that Vietnamese can be divided into three main dialects: northern dialect corresponding to Tonkin, central dialect corresponding to areas from Thanh Hoa province to Hai Van pass, southern dialect corresponding to areas from Hai Van pass to southern provinces [11]. In any case, this division is only relative because the geographical boundaries to divide the dialects are not completely clear. In fact, for the same regions, dialect can vary from a village to another. For three principal dialects above, in addition to the significant differences in vocabulary, it makes the listener easily perceive, distinguish between the dialects that is pronunciation modality. Phonetics of three main dialects differs significantly. For Vietnamese tone system, northern dialect has full six tones including level tone ("*thanh ngang*"), low-falling tone ("*thanh huyền*"), asking tone ("*thanh hỏi*"), rising tone ("*thanh sắc*"), broken tone ("*thanh ngã*") and heavy tone ("*thanh nặng*"), while central dialect has only five tones. For Thanh Hoa, Quang Binh, Quang Tri, Thua Thien voices and southern voice in general, there is no distinction between asking tone and broken tone. For Nghe An and Ha Tinh voices, broken tone and heavy tone are the same. In terms of prosody, three main dialects are entirely different.

The number of different Vietnamese dialects is very big. Traditionally, Vietnam is divided geographically into three regions: North, Centre, and South. The dialects for these three regions are also different both local vocabularies and pronunciation modalities. That is why we have chosen three representative dialects for these regions.

In our research, it is the difference between pronunciation modalities and but not local vocabulary that is exploited to identify three main dialects.

3. Building Vietnamese dialect corpus

3.1 Method for building Vietnamese dialect corpus

There are already dialectal corpora for some languages such as English [7], Chinese [8], Arabic [12], Thai [9], Hindi [3], [10]... For English, FRED is really a big dialect corpus which covers 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 permits the investigation of phenomena of non-standard morphosyntax beside analyses of phonetic or phonological details.

For Chinese, there are eight major dialectal regions. The authors in [8] 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 relatively large range of topics, text contents ensure tonal, 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 the 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. Signal to noise ratio of recorded corpus has been

evaluated based on spectrum subtraction method. Results showed that the average value of this ratio is 35 dB. This value is perfectly suited for the dialects identification system and voice recognition.

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.

3.2 The results of building Vietnamese dialect corpus VDSPEC

The number of different Vietnamese dialects is very big. Traditionally, Vietnam is divided geographically into three regions: North, Centre, and South. In fact, for the same regions, dialect can vary from a village to another. The dialects for these three regions are also different both local vocabularies and pronunciation modalities. That is why we have chosen three representative dialects for these regions. In our research, it is the difference between pronunciation modalities and but not local vocabulary that is exploited to identify three main dialects.

For VDSPEC corpus, the sampling frequency is 16000 Hz and 16 bits per sample. The speaker's average age is 21. At this age, voice quality is steady with full features for the local voice. Each dialect has 50 speakers including 25 men and 25 women. Hanoi voice is chosen for northern dialect, Hue voice for central dialect and Ho Chi Minh City voice for southern dialect. For each topic, the speaker reads 25 sentences and a sentence length is about 10 seconds. The total recording duration is 45.12 hours with the volume 4.84 GB.

4. Analysis of F0 variation

As above mention, fundamental frequency plays an important role in Vietnamese. Especially, the fundamental frequency variations of three dialects North, Center and South are different for the same tone. That is why we will analyse the variations of fundamental frequency for three dialects to describe these distinctions.

Three representatives male voices and three female voices were chosen to evaluate F0 variations. In fact, the duration of the tones is usually different. To make the difference more evident, these durations have been normalized by the same time length 0.5 seconds. F0 analysis results are shown in Fig. 1 and Fig. 2. Praat [13] was used to evaluate the fundamental frequency variation of Vietnamese tone in VDSPEC.

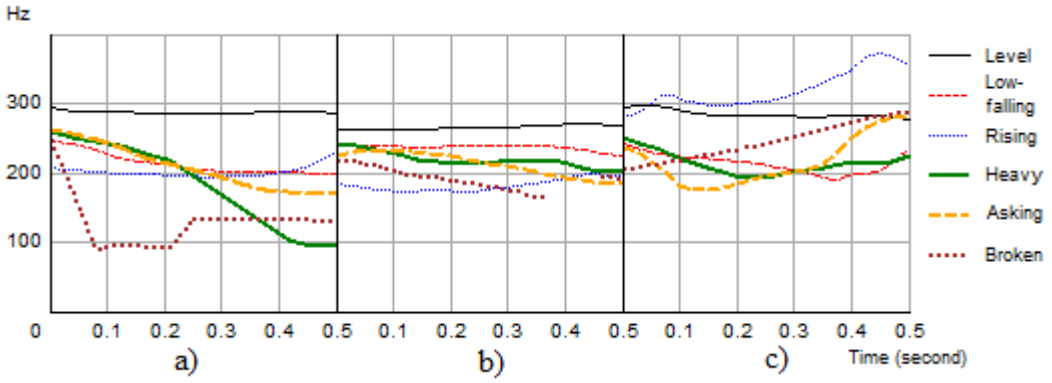

Fig. 1 F0 variation for six tones of female voices Hanoi (a), Hue (b), and HCM (c).

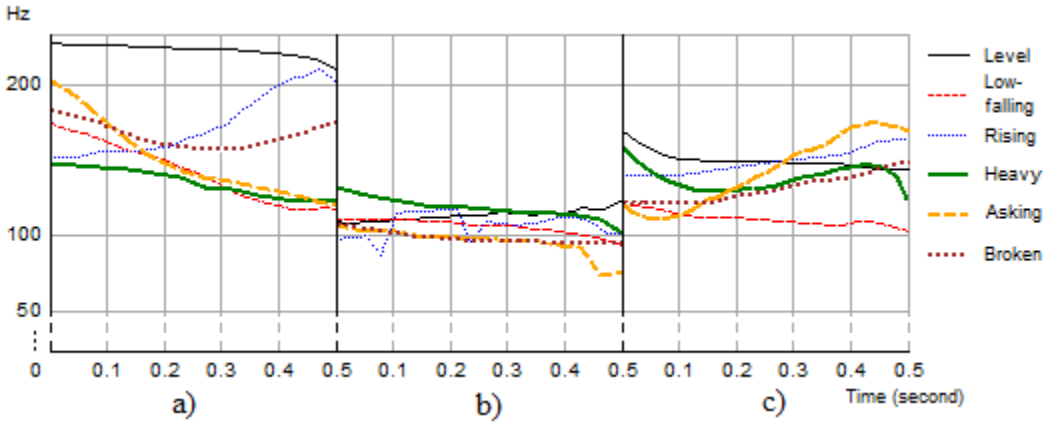

Fig. 2 F0 variation for six tones of male voices Hanoi (a), Hue (b), and HCM (c).

As we can see from figures 1 and 2, for level tone, F0 variation is rather small at around the mid level for three 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 and Ho Chi Minh City 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. In general, F0 of tones for Hue voices has the tendency to go down monotonously as low-falling or heavy tones for Hanoi voice or Ho Chi Minh City voice. For the broken tone of Ho Chi Minh City voice, the variation of F0 has a tendency to goes up at the end as asking tone of Hanoi voice. The range of F0 variation for 6 tones of Hue voice is smaller than Hanoi voice and Ho Chi Minh City voice.

The variation of F0 values for 150 speakers including 75 males and 75 females is also evaluated and is depicted by boxplots in figures from 3 to 8. These figures show F0 variation for male voices Hanoi (Hn-M), Hue (Hue-M), Chi Minh City (HCM-M), and female voices Hanoi (Hn-F), Hue (Hue-F), and Ho Chi Minh City (HCM-F). For each dialect, the number of female voices is 25 and the same for the number of male voices

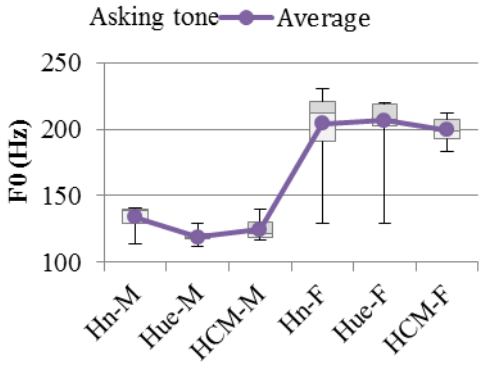

Fig. 3 F0 variation of asking tone.

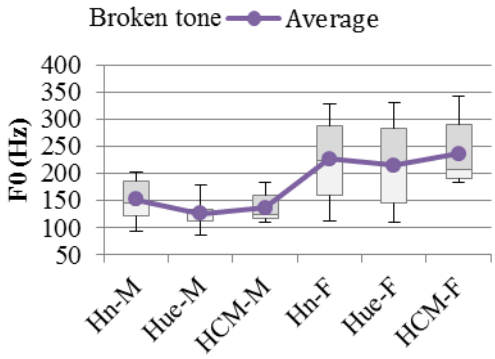

Fig. 4 F0 variation of broken tone.

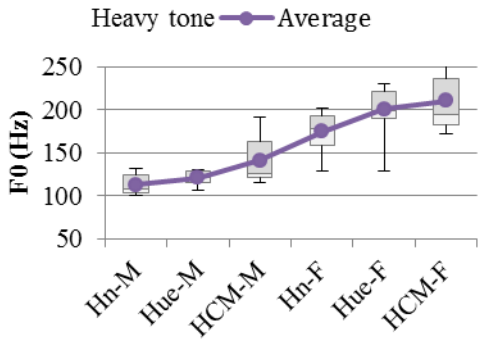

Fig. 5 F0 variation of heavy tone.

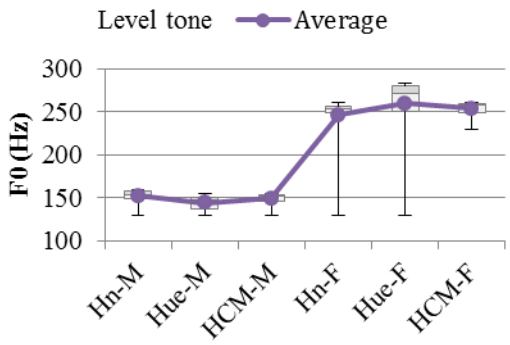

Fig. 6 F0 variation of level tone.

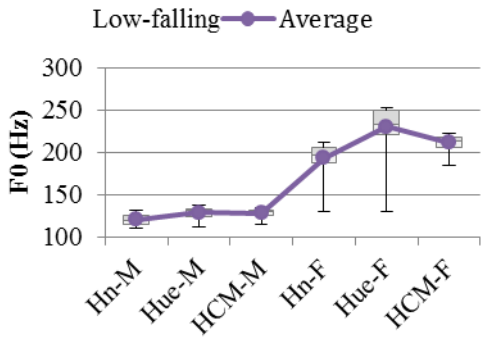

Fig. 7 F0 variation of low-falling tone.

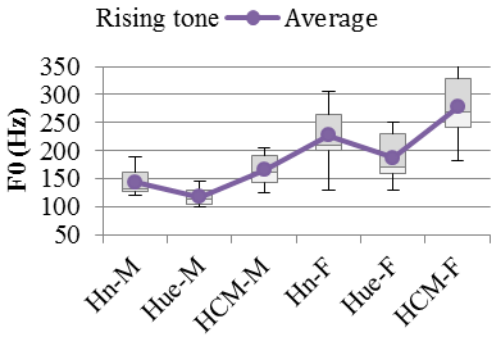

Fig. 8 F0 variation of rising tone.

From Fig. 3, the range of F0 variation for asking tone of Hue voices is smaller than the case of Hanoi voices and Ho Chi Minh City voices. For the level tone of Hue voices, the F0 variation range is larger than Hanoi voices and Ho Chi Minh City voices (Fig. 6).

For broken and rising tones, F0 of Hue voices tends to go down lower in comparison with Hanoi voices as in Fig. 4 and Fig. 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 Fig. 5 and Fig. 7. For heavy tone, the range of F0 variation of Ho Chi Minh City is larger than two other dialects.

Generally speaking, the direction and the range of F0 variation for Hue tones tends to be opposed to Hanoi tones. Except broken tone, the direction of F0 variation for Ho Chi Minh City voices are closed to Hanoi voice. For broken tone, the F0 variation of Ho Chi Minh City voices tends to go up as asking tone of Hanoi voices.

This conclusion is also consistent with the perception of reality about the difference between the pronunciation modality for the tones of three dialects

5. Preliminary classification of dialects using LDA

For the words with the ends such as "*nh*" or "*ch*" for example "*chính*" (“*main*”), "*kinh*" (“*fear*”), "*mạnh*" (“*strong*”), "*thành*" (“*wall*”), "*tỉnh*" (“*province*”), "*vĩnh*" (“*eternity*”), "*tích*" (“*product*”), "*tịch*" (“*confiscation*”) …, the pronunciation modalities of Hue voices and Ho Chi Minh City voices are different in comparison with Hanoi voices.

For Hue voices and Ho Chi Minh City voices, the nasalization level is stronger than Hanoi voices. In the case of Hue voices and Ho Chi Minh voices, the sound "*tích*" is pronounced as "*tứt*" (nonsense for Hanoi dialects), "*thành*" as "*thằn*" (nonsense for Hanoi dialects)... For the following presentation, we try to use spectral characteristics through MFCC (Mel Frequency Cepstral Coefficients) and F0 for the classification of three dialects using some above-mentioned sounds. LDA projection has been used to classify three dialects in our experiments.

In this case, LDA is based on projection matrix with the following equation

$$y = K.x \tag{1}$$

where $\boldsymbol{K}$ is the transformation matrix, $\boldsymbol{x}$ is the feature vector of the quantity to be classified:

$$x = \left[x_1 x_2 \ldots x_n\right]^T \tag{2}$$

and $\boldsymbol{y}$ is characteristic vector:

$$y = \left[y_1 y_2 \ldots y_m\right]^T \tag{3}$$

here ($m < n$).

For our dialect classification, $\boldsymbol{x}$ is the feature vector of the speech signal with n = 16, including 15 MFCC coefficients and one F0 value for each frame.

The objective of the LDA is to minimize the distance between the vectors of the same class and to maximize the distance between the class centers. These distances are represented by matrix $\boldsymbol{SW}$ and $\boldsymbol{SB}$. The rows of the matrix K are eigenvectors of matrix $\boldsymbol{\Sigma_{LDA}}$.

At first, the matrix $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$ and $\boldsymbol{\mu_0}$ are evaluated according to the following formula:

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_i \tag{4}$$

$$\mu_0 = \frac{1}{N_e}\sum_{i=1}^{N_e} \mu_k \tag{5}$$

$$\sum_k = \frac{1}{N_k}\sum_{i=1}^{N_k} (x_i - \mu_k)(x_i - \mu_k)^T \tag{6}$$

Next, matrix $\boldsymbol{\Sigma_{LDA}}$ is calculated as the following formula:

$$S_B = \frac{1}{N_e}\sum_{k=1}^{N_e} (\mu_k - \mu_0)(\mu_k - \mu_0)^T \tag{7}$$

$$S_W = \frac{1}{N_e}\sum_{k=1}^{N_e} \Sigma_k \tag{8}$$

$$\Sigma_{LDA} = \left(s_w + s_B\right)^{-1} \tag{9}$$

After that, the analysis of eigenvalue λ and eigenvector v_l for matrix $\boldsymbol{\Sigma_{LDA}}$ using the following equation:

$$\Sigma_{LDA} v_l = \lambda v_l \tag{10}$$

Finally, the vector projection is performed with $\boldsymbol{K}$:

$$K = \begin{bmatrix} v_1^T \\ \vdots \\ v_m^T \end{bmatrix} \tag{11}$$

$$y = Kx \tag{12}$$

Note that the number of non-zero eigenvalues of matrix $\boldsymbol{\Sigma_{LDA}}$ is the number of classes minus 1. All eigenvalues with values between 0 and 1 are indicators for distinction capability on the axes representing correspondent eigenvectors.

The figures from 9 to 14 are the classification results using LDA for the words "*chính*", "*kinh*", "*mạnh*", "*thành*", "*tỉnh*", "*vĩnh*" respectively. The results show that these words are classified clearly by using only MFCC and F0.

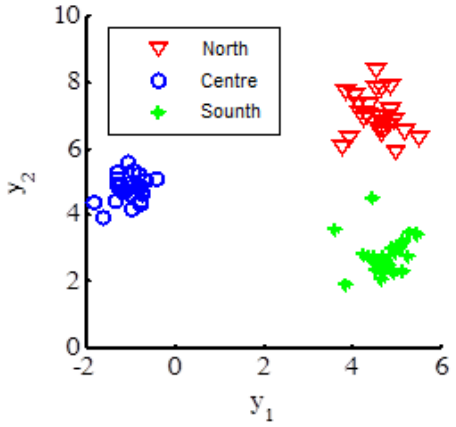

Fig. 9 Dialect classification using LDA for asking tone (word "*chính*").

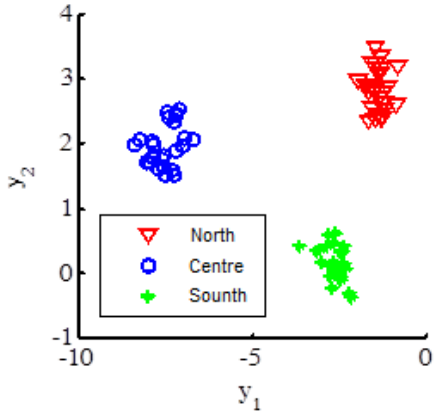

Fig. 10 Dialect classification using LDA for level tone (word "*kinh*").

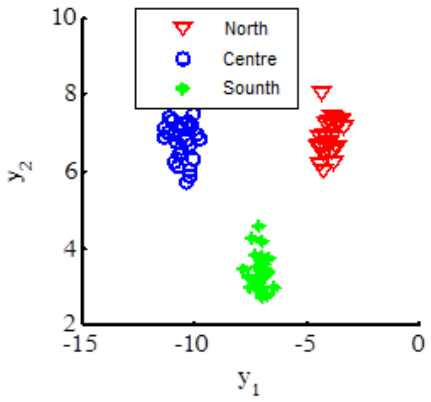

Fig. 11 Dialect classification using LDA for heavy tone (word "*mạnh*").

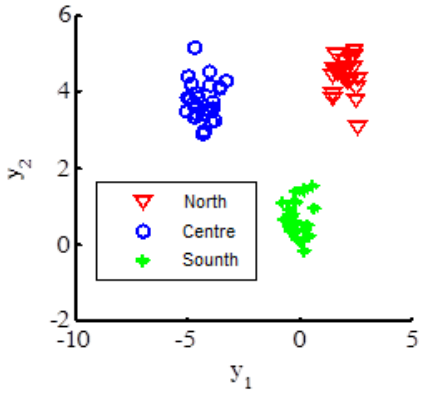

Fig. 12 Dialect classification using LDA for low-falling tone (word "*thành*").

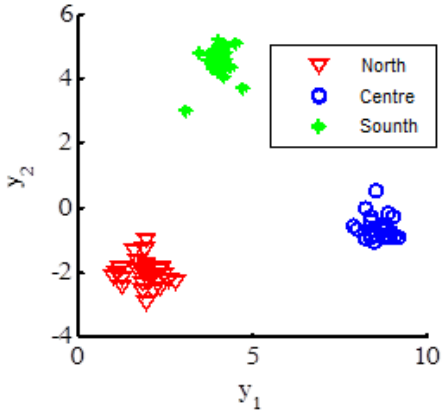

Fig. 13 Dialect classification using LDA for asking tone (word "*tỉnh*").

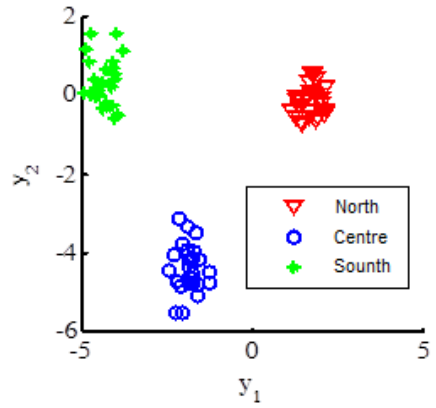

Fig. 14 Dialect classification using LDA for broken tone (word "*vĩnh*").

6. Dialect identification using GMM with formants, their bandwidths and F0 parameters

Normally, formant frequencies and bandwidths are vocal tract parameters. The formants are frequencies of vocal tract resonances. The first two formants are the most important because they decide the speech quality [14]. Formants and their bandwidths have been used for a lot of research on speech processing such as accent identification [15], [16], [17], speech recognition [18], speaker identification [19], study on genders and ethnical accents [20], [21], [22], dialect identification [10], [23], [24], [25].

In our experiments, the values of the first four formants and their bandwidths are calculated using Praat. These values are combined with F0 and its variants. The experiments are performed using the baseline of Gaussian component number for which the Gaussian component number equals 20. The dialectal identification results with different combinations of these parameters are presented in Table 1.

Beside *F0* value, some quantities derived from *F0* are calculated as follows

- The derivative *F0* (*diffF0(t)*):

$$diffF0(t) = dF0(t)/dt \tag{13}$$

- The trend upward or downward of *F0* for each sentence (*cdF0(t)*):

$$cdF0(t) = \begin{cases} -1 & if\ ((F0_i - F0_{i-1}) \le -3) \\ 0 & if\ (-3 < (F0_i - F0_{i-1}) < 3) \\ 1 & if\ ((F0_i - F0_{i-1}) \ge 3) \end{cases} \tag{14}$$

- The normalized *F0* according to average *F0* for each sentence (*F0sbM(t)*):

$$F0sbM(t) = F_0(t) / \overline{F_0(t)} \tag{15}$$

- The normalized *F0* according to average and standard deviation *F0* (*F0sbMSD*(t)):

$$F0sbMSD(t) = \frac{F_0(t) - \overline{F_0(t)}}{\sigma F_0(t)} \tag{16}$$

- The derivative *LogF0* (*diffLogF0(t)*):

$$diffLogF0(t) = d\ LogF0(t)/dt \tag{17}$$

- The normalized *LogF0* according to min *LogF0* and max *LogF0* for each sentence (*LogF0sbMM(t)*):

$$LogF0sbMM(t) = \frac{LogF_0(t) - \min LogF_0(t)}{\max LogF_0(t) - \min LogF_0(t)} \tag{18}$$

Table 1: Recognition results using formants, corresponding bandwidths and F0 parameters.

Index	Formants+Bandwidths +F0 Parameters	Recognition Rate
1	Formants+Bandwidths	66.3%
2	*F0*	67.5%
3	*diffF0(t)*	65.2%
4	*cdF0(t)*	67.0%
5	*F0sbMM(t)*	67.8%
6	*F0sbM(t)*	64.3%
7	***F0sbMSD(t)***	**72.2%**
8	*LogF0(t)*	71.6%
9	*diffLogF0(t)*	66.8%
10	*LogF0sbMM(t)*	67.7%
11	*LogF0sbM(t)*	68.7%
12	*LogF0sbMSD(t)*	66.8%

- The normalized *LogF0* according to average *LogF0* for each sentence (*LogF0sbM(t)*):

$$LogF0sbM(t) = logF_0(t) / \overline{logF_0(t)} \quad (19)$$

- The normalized *LogF0* according to average and standard deviation *LogF0* (*LogF0MSD(t)*):

$$LogF0MSD(t) = \frac{logF_0(t) - \overline{logF_0(t)}}{\sigma logF_0(t)} \quad (20)$$

The highest recognition rate is 72.2% for the case using formants, their bandwidths and the normalized *F0* according to average and standard deviation *F0* (*F0sbMDS(t)*).

7. Effect of Gaussian component number on dialect recognition performance

For this experiment, formants, their bandwidths + *F0sbMSD(t)* are chosen and the Gaussian component number *M* is taken from 16 to 4096. GMM was trained and evaluated with this range of components. The DET (Detection Error Tradeoff) curves for different values of Gaussian component number are depicted in Fig. 15.

From Fig. 15, generally, at first, the increase in *M* decreases the dialect recognition performance but increases the dialect recognition performance after that. The maximum recognition rate is 72.2% when *M* equals 20. In Fig. 15, the points indicated by °'s are weighted averages of the missed detection and false alarm rates or the minimum values of the Detection Cost Function (*DCF*). These values are calculated as the following [26]:

$$DCF = C_{miss}.P_{miss}.P_{true} + C_{fa}.P_{fa}.P_{false} \quad (21)$$

where C_{miss} is the cost of a miss (rejection), C_{fa} is the cost of an alarm (acceptance), P_{true} is the *a priori* probability of the target, P_{fa} is the false alarm probability and P_{false} = 1 - P_{true}. C_{miss} = C_{fa} = 1. The minimum value of the *DCF* for *M*=20 corresponds to the point which is closest to the origin.

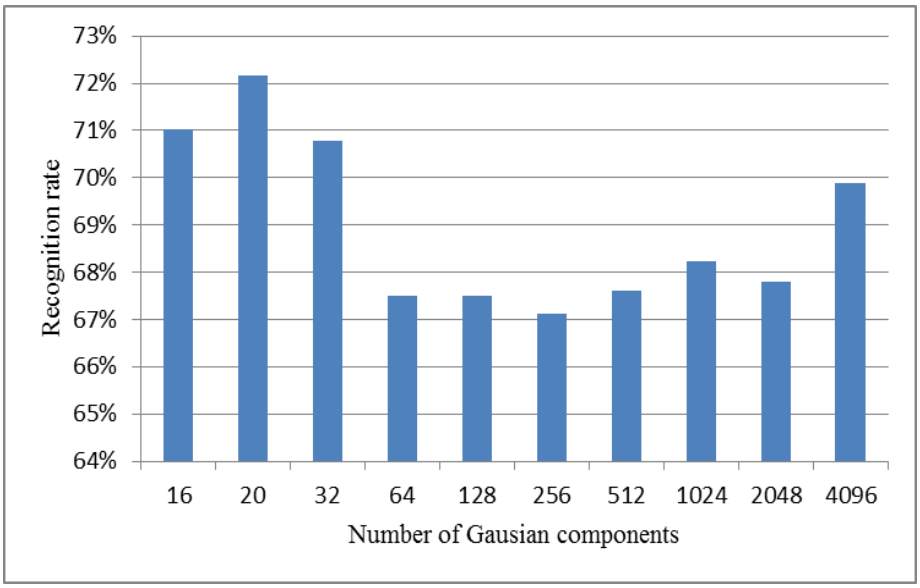

Fig. 14 Recognition performance in function of mixture number.

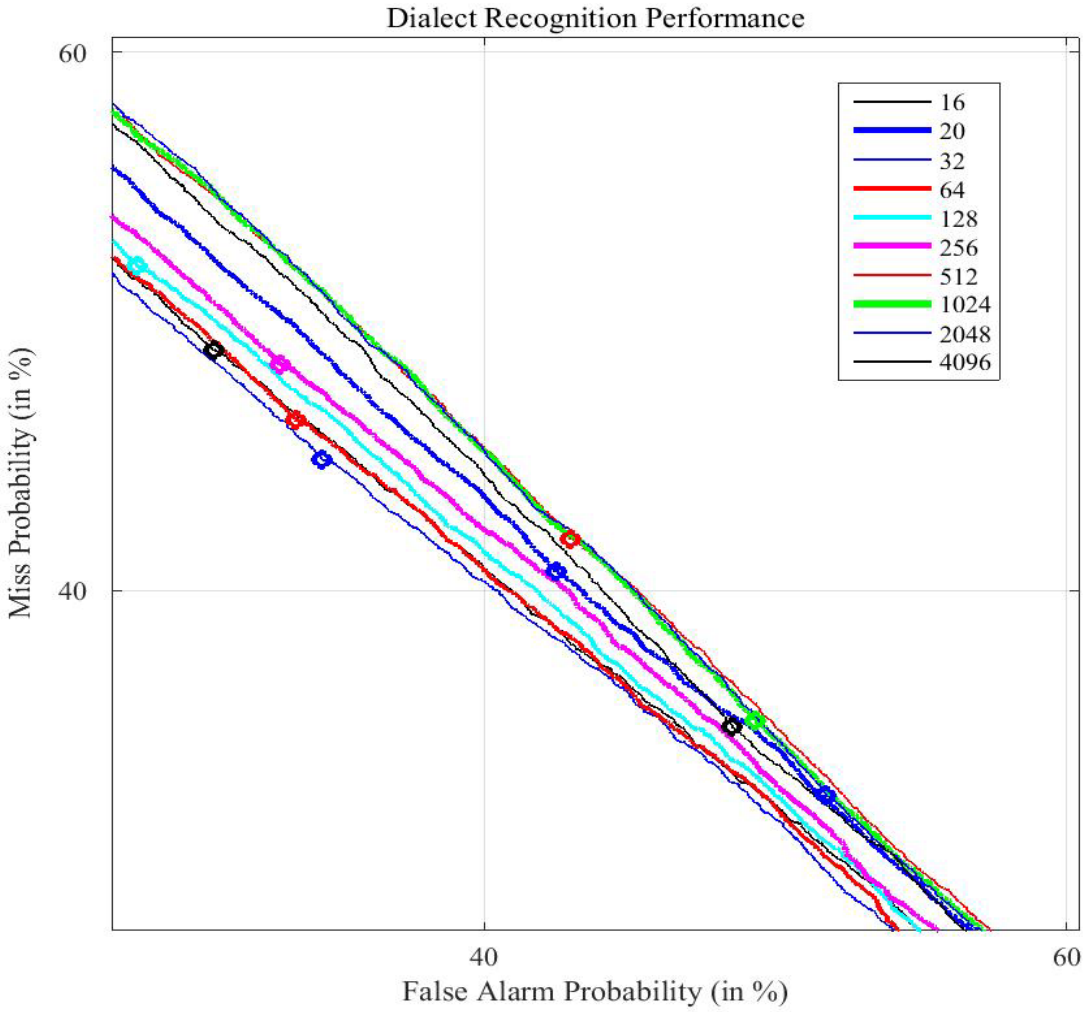

Fig. 15 DET curves with Gaussian component number from 16 to 4096.

8. Conclusions

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 results of *F0* variation and classification using LDA projections showed that there is a very good distinction between pronunciation modalities of three dialects Hanoi, Hue and Ho Chi Minh City.

This corpus is not only useful for the study of dialects recognition and speech recognition but also suitable for the study of Vietnamese synthesis. The recognition rate is highest in the case using formants, their bandwidths and normalized *F0* according to average and standard deviation *F0*. The best recognition rate is obtained with 20 Gaussian components. These research results can continue to develop for its application in the automatic recognition systems to enhance Vietnamese recognition performance.

References

[1] 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).

[2] Hirayama N., Yoshino K., Itoyama K., Mori S., Okuno, H.G, "Automatic Speech Recognition for Mixed Dialect Utterances by Mixing Dialect Language Models." *Audio, Speech, and Language Processing, IEEE/ACM Transactions on* (Volume:23 , Issue: 2), pp 373 - 382, Feb (2015).

[3] Shweta Sinha, "Analysis and Recognition of Dialects of Hindi Speech." *International Journal of Scientific Research in Multidisciplinary Studies*, Volume-1, Issue-1, August (2015), pp 26-33.

[4] 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

[5] T.T. Vu, D.T. Nguyen, M.C. Luong, and J-P. Hosom, "Vietnamese Large Vocabulary Continuous Speech Recognition." *In INTERSPEECH*, Lisbon, Portugal, September, (2005).

[6] Vu, Q., Demuynck, K., Compernolle, D.V, "Vietnamese Automatic Speech Recognition: the FlaVoR Approach." *ISCSLP 2006*, Kent Ridge, Singapore (2006).

[7] *Bernd Kortmann, A Comparative Grammar of British English Dialects*, 2005, Walter de Gruyter.

[8] Jing Li et al., "A Dialectal Chinese Speech Recognition Framework", *Journal of Compute. Sci. & Technol*., Jan.2006, Vol. 21, No. 1, pp. 106-115.

[9] 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 .

[10] Shweta Sinha, Aruna Jain, S. S. Agrawal, "Acoustic-Phonetic Feature Based Dialect Identification in Hindi Speech", *International Journal on Smart Sensing and Intelligent Systems*, Vol. 8, No. 1, March 2015, pp. 235-254.

[11] Hoàng Thị Châu, *Phương ngữ học tiếng Việt*. NXB Đại học Quốc gia Hà Nội (2009).

[12] Fadi Biadsy, Julia Hirschberg, *Using Prosody and Phonotactics in Arabic Dialect*. Identification. Interspeech, Vol. 1, pp 208-211 (2009).

[13] www.praat.org

[14] Carlson, Rolf, Gunnar Fant, and Björn Granström, "Two-formant models, pitch, and vowel perception." *Acta Acustica united with Acustica* 31.6, 1974, pp. 360-362.

[15] Stantic, Dejan, and Jun Jo. "Accent Identification by Clustering and Scoring Formants." *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* 6.3, 2012, pp. 379-384.

[16] Hanani, Abualsoud, Martin J. Russell, and Michael J. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech." *Computer Speech & Language* 27.1, 2013, pp. 59-74.

[17] Mannepalli, Kasiprasad, P. Nrahari Sastry, and V. Rajesh, "Accent detection of Telugu speech using prosodic and formant features." *Signal Processing And Communication Engineering Systems (SPACES), 2015 International Conference on. IEEE*, 2015, pp. 318-322.

[18] Garner, Philip N., and Wendy J. Holmes, "On the robust incorporation of formant features into hidden Markov models for automatic speech recognition." *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on. Vol. 1. IEEE*, 1998, pp. 1-4.

[19] Becker, Timo, Michael Jessen, and Catalin Grigoras, "Forensic speaker verification using formant features and Gaussian mixture models." *Interspeech*, 2008, pp. 1505-1508.

[20] Yusnita, M. A., et al., "Acoustic analysis of formants across genders and ethnical accents in Malaysian English using ANOVA." *Procedia Engineering 64*, 2013, pp. 385-394.

[21] Gelfer, Marylou Pausewang, and Victoria A. Mikos. "The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels." *Journal of Voice* 19.4, 2005, pp. 544-554.

[22] Hillenbrand, James M., and Michael J. Clark,

"The role of f 0 and formant frequencies in distinguishing the voices of men and women." *Attention, Perception, & Psychophysics* 71.5, 2009, pp. 1150-1166.

[23] Hagiwara, Robert, "Dialect variation and formant frequency: The American English vowels revisited." *The Journal of the Acoustical Society of America 102.1*, 1997, pp. 655-658.

[24] Jacewicz, Ewa, and Robert Allen Fox, "The effects of dialect variation on speech intelligibility in a multitalker background." *Applied Psycholinguistics 36.03*, 2015, pp. 729-746.

[25] Fox, Robert Allen, and Ewa Jacewicz, "Cross-dialectal variation in formant dynamics of American English vowels." *The Journal of the Acoustical Society of America 126.5*, 2009, pp. 2603-2618.

[26] Martin, Alvin, et al., "The DET curve in assessment of detection task performance." *National Inst. Of Standards and Technology Gaithersburg Md*, 1997.

Pham Ngoc Hung is a lecturer at Faculty of Information Technology in Hungyen University of Technology and Education. He received his Bachelor (2005) in Faculty of Informatics Technology - Le Quy Don Technical University and Master degree (2010) in Faculty of Informatics Technology – Hanoi VNU University of Engineering and Technology (Hanoi VNU-UET). His research interests include speech recognition and embedded system.

Trinh Van Loan is an Associate Professor of Informatics at the School of Information and Communication Technology, Hanoi University of Science and Technology. He received his doctorate in Electronics Systems from Grenoble Institute of Technology, France. His present research and teaching interests are in the fields of Speech Processing, Digital Signal Processing, and Embedded Systems.

Nguyen Hong Quang is a lecturer at School of Information and Communication Technology in Hanoi University of Science and Technology. He received his Bachelor of Science (2000) of Information Technology in Hanoi University of Science and Technology, Master degree (2004) in Information Processing and Communication in Hanoi University of Science and Technology and Doctor of Information Technology (2008) in University of Avignon, Avignon, France. His research interests include signal processing, machine learning, speech processing, speech recognition, especially for Vietnamese.