Journal of Computer Science and Cybernetics, V.30, N.1 (2014), 28–38

VIETNAMESE RECOGNITION USING TONAL PHONEME BASED ON MULTI SPACE DISTRIBUTION

NGUYEN VAN HUY[1], LUONG CHI MAI[2], VU TAT THANG[2], DO QUOC TRUONG[3]

[1]*Electronic faculty, Thai Nguyen University of Technology, VietNam*

[2]*Institute of Information Technology, Vietnam Academy of Sience and Technology, Vietnam*

[3]*Graduate School of Information Science, Nara Institute of Science and Technology, Japan*

Tóm tắt. Báo cáo trình bày việc áp dụng mô hình Markov ẩn phân bố đa không gian Multi Space Distribution Hidden Markov Model (MSD-HMM) cho nhận dạng tiếng Việt. Nghiên cứu đề xuất một kiểu mô hình MSD-HMM để mô hình hoá cho các âm vị có chứa thông tin thanh điệu với đặc trưng đầu vào gồm bốn lớp độc lập. Các âm vị có thanh điệu được tạo ra bằng cách bổ sung thêm các ký hiệu thanh điệu tương ứng với từ chứa âm vị đó dựa theo bảng ngữ âm quốc tế (International Phonetic Alphabet). Kết quả nhận dạng sau khi áp dụng mô hình MSD-HMM trên tập âm vị có thanh điệu tốt hơn so với hệ thống cơ sở là 2.49%. Báo cáo cũng trình bày một cách tiếp cận để trích trọn đặc trưng thanh điệu nhằm tìm ra dạng đặc trưng thanh điệu phù hợp với mô hình MSD-HMM. Các kết quả thử nghiệm trong nghiên cứu này đã chỉ ra rằng mô hình MSD-HMM kết hợp với tập từ vị có thanh điệu đã làm tăng đáng kể độ chính xác nhận dạng, đồng thời cho thấy đặc trưng thanh điệu là một thành phần quan trọng trong các hệ thống nhận tiếng Việt.

Từ khóa. Phân bố đa không gian, nhận dạng tiếng Việt, đặc trưng thanh điệu, nhận dạng thanh điệu.

Abstract. This paper presents an approach of Multi Space Distribution Hidden Markov Model (MSD-HMM) for Vietnamese recognition. An MSD-HMM prototype with four independent streams is proposed for modeling the Vietnamese phonemes which embedded tonal information corresponding to its syllable. These phonemes are built by adding tonal symbol to each phoneme syllables based on the International Phonetic Alphabet (IPA). This approach improves 2.49% accuracy compared to the baseline system. A process of tonal feature extraction that is suitable for modeling by MSD-HMM is also described. The result shows that the performance of MSD-HMM and tonal phoneme is better than the baseline system, and the tonal phoneme and tonal feature are important components for Vietnamese recognition.

Key words. Multi space distribution, tone recognition, Vietnamese recognition, pitch feature.

1. INTRODUCTION

Vietnamese is a tonal monosyllable language in which each word has only one of six tones. There are probably six different meanings when combining a word with six different tones,

because of some combination of word and tone that means nothing. Therefore, a good automatic speech recognition (ASR) system for Vietnamese should also include tone recognition. The acoustic features widely known for ASR are Mel Frequency Spectral Coefficient (MFCC) and Perceptual Linear Prediction (PLP), but these features do not contain tonal feature which can represent tone information. The tonal feature can be obtained through the fundamental frequency F0 (or pitch feature). In fact, F0 is widely used for representing tonal feature in both ASR and speech synthesis. However, the problem is that F0 does not exist in the unvoiced region, so it cannot be presented by a continuous value as in the voice region. Consequently, F0 feature vector that is extracted from a speech sample would consist of discrete and continuous values. This is a difficulty for the ASR system based on Hidden Markov Model (HMM), because HMM only models discrete pattern or continuous pattern individually.

Vietnamese speech recognition integrated tone recognition for larger vocabulary continuous speech is only at the beginning phase of development. Recently, there are several results (see, e.g. [1–5]) proposed some approaches for tone recognition of Vietnamese, but these approaches model tones by applying a continuous tonal feature. The methods to extract tonal feature in those papers try to fix the errors in the unvoiced region or replace the unvoiced pattern by a random continuous value. In this paper, we present another approach for Vietnamese recognition integrated tone recognition based on MSD-HMM by applying tonal phonemes. This approach models tonal phonemes by using a combination of tonal feature and acoustic feature, but the tonal feature could contain both continuous and discrete values and it do not need any method to fix the non-existence of $F0$ in the unvoiced regions.

This paper is organized as follows. In section 2, the basic and a prototype of MSD-HMM applying for Vietnamese are described. In section 3, we present the phonetic structure of Vietnamese, and propose a set of Vietnamese tonal phonemes that is appropriate for the MSD-HMM model. The process of tonal feature extraction is presented in section 4. The experiments and the results are given in section 5. We conclude the paper in section 6 with the summary of this study.

2. BASIC OF MULTI SPACE DISTRIBUTION

Hidden Markov Model (HMM) is widely used for automatic speech recognition, but HMM is defined only for modeling discrete pattern or continuous pattern individually. Therefore, the difficulty on HMM-based pitch modeling is that a raw pitch feature would consist of both discrete pattern for the unvoiced region and continuous pattern for the voice region, since pitch only exists on the voice region. In general, there are two approaches to solution of this problem. The first approach replaces unvoiced patterns by heuristic values, and then models these patterns by using the continuous HMM. The second approach adapts HMM to model pitch feature which could contain both discrete and continuous patterns. Multi Space Distribution (MSD) was proposed by Tokuda which belongs to the second approach. MSD is defined to model the pitch [6][7] without any heuristic information and it was successfully applied for Mandarin [8]. It can model the feature that consists of both continuous and discrete values, so we do not need using any method for interpolation of artificial values into the unvoiced regions of pitch.

Multi Space Distribution Hidden Markov Model (MSD-HMM) is proposed based on MSD, which is similar to the original HMM model. There is only one difference on observation probability function. MSD assumes that there is a space $\Omega = \bigcup_{g}^{G} \Omega_g$ which consists of G

subspaces, where Ω_g is a subspace of n_g dimensionals. The feature is that n_g can be different in different subspaces and can be zero. If $n_g = 0,\ x$ will represent a discrete value, otherwise x is a continuous value for all $n_g > 0$. Each subspace Ω_g has a weight ω_g to present its prior probability in Ω, where $\sum_{g}^{G} \omega_g = 1$. Then an observation vector o consists of two elements: $o = \{x, l\}$, where x is a random variable, and I is a set of space indexes for specifying the space that x belongs to. The observation probability function of vector x in the normal HMM is defined by Equation 1, then it is defined by Equation 2 in MSD-HMM model.

$$b_i(x) = \mathcal{N}_i(x), \tag{1}$$

$$b_i(o) = \sum_{g \in I} \omega_{ig} \mathcal{N}_{ig}(x), \tag{2}$$

where, $o = \{x, I\}$, $x \in R^{n_g}$, i is i^{th} state of HMM model, g is g^{th} subspace of Ω, $\mathcal{N}_{ig}(x)$ and $\mathcal{N}_g(x)$ are the probability density functions (pdf) of random variable vector x. $\mathcal{N}_g(x)$ is undefined for $n_g = 0$ with normal HMM, but MSD-HMM defined by $\mathcal{N}_{ig}(x) = 1$. Therefore, $b_i(o)$ can be calculated for both cases of discrete and continuous values.

The output observation probability function is defined by (2). An N-state MSD-HMM λ is specified by initial state probability distribution set $\pi = \{\pi_j\}_{j=1}^{N}$, the state transition probability distribution set $A = \{a_{ij}\}_{i,j=1}^{N}$ (where a_{ij} is the probability for state sitransits to state s_j), and state output probability distribution set $B = \{b_i(o)\}_{i=1}^{N}$. Given an observation sequence $O = \{o_1, o_2, o_3, ..., o_T\}$, the observation probability of O is defined by

$$P(O|\lambda) = \sum_{q,l} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) = \sum_{q,l} \prod_{t=1}^{T} a_{q_{t-1}q_t} w_{q_t l_t} \mathcal{N}_{q_t l_t}(x)$$

where $q = \{q_1, q_2, ..., q_T\}$ is a possible states sequence and $l = \{l_1, l_2, ..., l_T\}$ is a possible indices sequence corresponding to observation sequence O. The parameters of λ model are also estimated by the forward and backward algorithms as the normal HMM model.

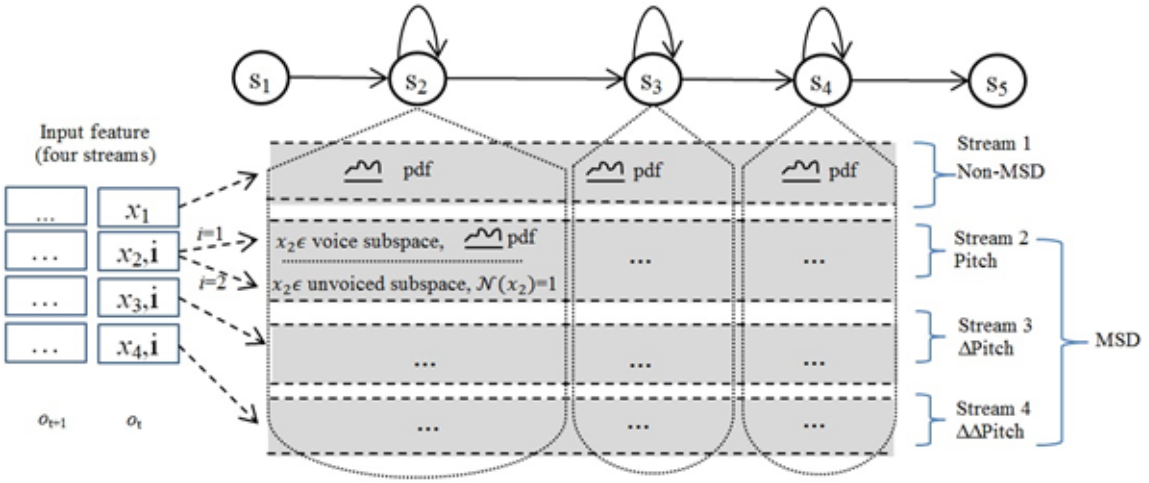

Figure 1. 5-states MSD-HMM prototype with four independent streams input feature

In the context of pitch modeling by using MSD-HMM defined above, the pitch feature can contain both discrete and continuous values. In this paper, we apply two subspaces $\Omega = \{\Omega_{n_1}, \Omega_{n_2}\}$ corresponding to voice and unvoiced subspaces, where $n_1 = 0$ and $n_2 = 1$. An observation vector o consists of two elements $o = \{x, i\}$. If x is a continuous value then i is

set to 1 for specifying the case x belongs to the voice subspace. If x is a discrete value then i will be set to 2 for specifying the case x belongs to the unvoiced subspace. These values of x and i are determined at the pitch extraction phase. In order to apply MSD for Vietnamese, we propose a left-right MSD-HMM prototype of 5 states to model input feature which has four independent streams. The first stream can be an acoustic feature or a combination of acoustic feature and continuous pitch feature, and this stream is modeled by the normal HMM. The 2nd, 3rd, and 4th streams contain pitch, delta of pitch and double delta of pitch in that order. The feature in these streams can consist of both continuous and discrete values, and they are modeled by MSD. Figure 1 shows this prototype.

3. TONAL PHONEME FOR VIETNAMESE

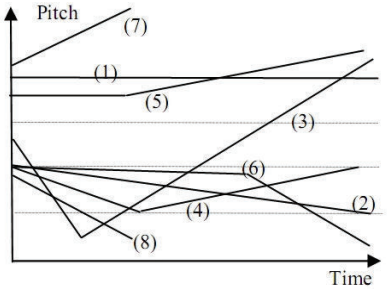

Figure 2. Vietnamese Tone Patterns

Tone			
Initial	Final		
	Onset	Nucleus	Coda

Table 1. Structure of Vietnamese syllable

Vietnamese is a tonal monosyllable language, each syllable may be considered as a combination of Initial, Final and Tone components in Table 1. The Initial component is always a consonant, or it may be omitted in some syllables (or seen as zero Initial). There are 21 Initials and 155 Final components in Vietnamese. The total of pronounceable distinct syllables in Vietnamese is 18958, but the used syllables in practice are only around 7000 different syllables [9]. The Final can be decomposed into Onset, Nucleus and Coda. The Onset and Coda are optional and may not exist in a syllable. The Nucleus consists of a vowel or a diphthong, and the Coda is a consonant or a semi-vowel. There are 1 Onset, 16 Nuclei and 8 Codas in Vietnamese. There are six lexical tones in Vietnamese, and they can affect word meaning. They are called high (or mid) level, low falling, dipping-rising, creaking-rising, high (or mid) rising, constricted correspond with Figure 2 (from 1 to 6) [10]. These six different tones applied to a syllable could result in six distinct words. Syllables with a closure coda can only go with rising tones and drop tones [11][12]. As Figure 2 (7 and 8), 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 [13], [14]. Therefore, most linguists who study Vietnamese acoustics claim that the Vietnamese language contains 8 different tones base on F0 contours, as show in Figure 2.

In [1], we proposed three kinds of phoneme set for Vietnamese recognition system using input features included pitch, and we obtained the best result on phoneme set which embedded tone information. In [5], we conducted experiments using this approach for Vietnamese and Cantonese on telephone speech corpus, and it gives about 1% improvement compared to phoneme set without tone information. Following this idea, we build two kinds of phonemic set for testing MSD corresponding to the phonemic structure as Table 1. The first set (PS1) have 44 phonemes which are created based on IPA without any tonal information. The second

set (PS2) is a modification of PS1. Every Nucleus phoneme and Coda phoneme in the final part of each syllable are combined with a tonal symbol according to its syllable, which is so-called tonal phoneme, the initial elements are the same as PS1. By this way, the number of phonemes is up to 153. Table 2 presents some examples that describes the approach to build these phoneme sets. The proposed phonemes of PS1 and PS2 in this paper are shown in Table 3.

4. TONAL FEATURE

There are some methods well-known to extract pitch feature. In this experiment, we apply two methods widely used for extraction pitch feature. They are Average Magnitude Difference Function (AMDF) [15] and Normalized Cross-Correlation (NCC) [16]. Both of AMDF and NCC are modified versions of the basic Auto-correlation. AMDF is defined by Equation 3 and NCC is defined by Equation 4.

$$D(\tau) = \frac{1}{N-\tau-1} \sum_{n=1}^{N-\tau-1} |x(n) - x(n+\tau)| \tag{3}$$

Table 2. Examples of creating tonal phoneme set based on the set without tonal information

English	Vietnamese	Telex	Tone	Phoneme Set 1 (PS1)	Phoneme Set 2 (PS2)
Zero	Không	Khoong	1	kh oo ngz	kh oo_ ngz_
Boat	Thuyền	Thuyeen**f**	2	th u iee nz	th u**f** iee**f** nz**f**
Act	Diễn	Dieen**x**	3	d iee nz	d iee**x** nz**x**
Seven	Bẩy	Baay**r**	4	b aa iz	b aa**r** iz**r**
Four	Bốn	Boon**s**	5	b oo nz	b oo**s** nz**s**
Spot	Mụn	Mun**j**	6	m u nz	m u**j** nz**j**
Style	Mốt	Moot**s**	7	m oo tc	m oo**s** tc**s**
One	Một	Moot**j**	8	m oo tc	m oo**j** tc**j**
Unit	Chiếc	Chieec**s**	7	ch iee c	ch iee**s** c**s**
Cheat	Bịp	Bip**j**	8	b i pc	b i**j** pc**j**

Table 3. The phonemes of PS1 and PS2

Phoneme Set	Initial	Onset	Nucleus/Coda
PS1	b d dd g h k kh l m n ng nh p ph r s t th tr v	w	a aa aw e ea ee i ie iz kc mz ng ngz nh nz o oa oo ow pc tc u uo uw uz wa
PS2	b d dd g h k kh l m n ng nh p ph r s t th tr v	w	a_ aa_ aaf aaj aar aas aax af aj ar as aw_ awf awj awr aws awx ax e_ ea_ eaf eaj ear eas eax ee_ eef eej eer ees eex ef ej er es ex i_ ie_ ief iej ier ies iex if ij ir is ix iz_ izf izj izr izs izx kcj kcs mz_ mzf mzj mzr mzs mzx ngz_ ngzf ngzj ngzr ngzs ngzx nz_ nzf nzj nzr nzs nzx o_ oa_ oaf oaj oar oas oax of oj oo_ oof ooj oor oos oox or os ow_ owf owj owr ows owx ox pcj pcs tcj tcs u_ uf uj uo uof uoj uor uos uox ur us uw_ uwf uwj uwr uws uwx ux uz_ uzf uzj uzr uzs uzx wa_ waf waj war was wax

$$NCC(\tau) = \frac{1}{\sqrt{e_n e_\tau}} \sum_{n=0}^{N-\tau-1} x(n)x(n+\tau) \tag{4}$$

$$e_i = \sum_{n=i}^{i+N-1} x^2(n) \tag{5}$$

where $x(n)$ is the input speech sample, N is the length of the speech analysis window, τ is the lag number in range between 0 and $N-1$. The experiments of previous papers show that NCC detects pitch better than AMDF since the peaks are more prominent and less affected by the rapid variations in the signal amplitude. However, $NCC(\tau)$ is calculated by normalizing based on multiplication of energies for all n in the range of e_i as (5). Therefore, NCC would detect pitches not only in the voice regions, but also in the unvoiced regions. The untrue pitches in the unvoiced regions are the artificial values. In fact, the pitch output that is acquired by NCC contains a ratio of unvoiced values to voice values, is smaller than AMDF. Figure 4 shows an example for describing this problem. Figure 4 (a, d) shows original pitch contours after applying AMDF and NCC for a speech sample of utterance "chắc chữa được bách bệnh". The pitch contour of NCC output is almost continual on all frames of utterance, whereas the pitch contour of AMDF output is much faultier. But the contour of AMDF output is more approximate to the true pitch contour. For HMM-based ASR system, NCC can be used as a continuity pitch feature for modeling by the normal HMM directly, but AMDF is not. In this paper, we want to find out which kind of pitch feature is more effective for applying the MSD-HMM model.

For evaluating the performance of MSD-HMM compared to the normal HMM, we extract two kinds of pitch feature. Firstly, the original pitch features are obtained by applying AMDF or NCC on the input speech samples, we replace all of the values in unvoiced regions by zero values, and then we do normalizing to obtain the first kind of pitch feature which is assumed as a continuous pitch feature that so-called ToneF1 as Figure 3. The second kind so-called ToneF2 is shown in Figure 3. The values in unvoiced regions of the original pitch features are replaced by a specific symbol, and then we do normalizing only for the frames in the voice regions. All of frames in the unvoiced regions are skipped. The normalized values are shown in Equation 6 which were proposed and tested in our previous paper [2]. Thereafter, ToneF1 is modeled by the normal HMM for comparing to ToneF2 which is modeled by MSD-HMM. Figure 3 describes our approach for pitch extraction, Figure 4 (b, e) presents ToneF1 contours, Figure 4 (c, f) presents ToneF2 contours.

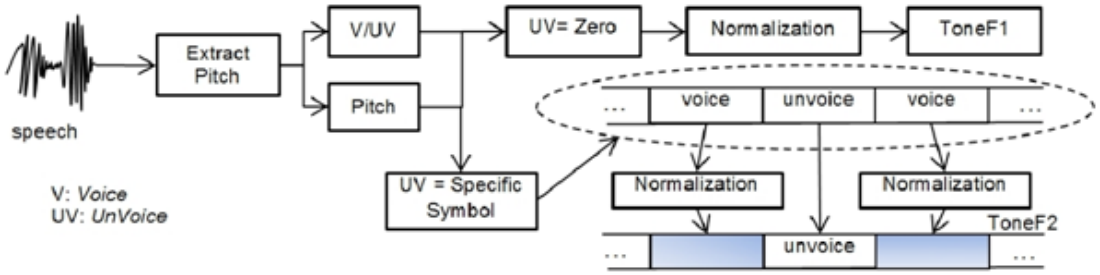

Figure 3. Block diagram of pitch feature extraction

$$f0_m = \frac{\log(F0_m) - mean(F0)}{Dev(F0)},\ m = 0, 1, ..., M, \tag{6}$$

$$Mean(F0) = \frac{1}{M} \sum_{m=0}^{M-1} F0_m, \tag{7}$$

$$Dev(F0) = \sqrt{\frac{1}{N}\sum_{m-0}^{M-1}(F0_m - Mean(F0))^2}, \quad (8)$$

where M is the number of frames of an input pitch $F0$, $F0_m$ is the m^{th} frame in the $F0$, $Mean(F_0)$ is the mean of F_0 that is computed as Equation 7, $Dev(F0)$ is the standard derivation of $F0$ that is computed as Equation 8.

5. EXPERIMENT SETUP

5.1. Speech corpus

The data used in our experiments is the Voice of Vietnam (VOV) data which is a collection of story reading, mailbag, new reports, and colloquy from the radio program the Voice of Vietnam. There are 23424 utterances in this corpus including about 30 male and female broadcasters and visitors. The number of distinct syllables with tone is 4923 and the number of distinct syllables without tone is 2101. The total time of this corpus is about 19 hours. We separate into training set of 17 hours and decoding set of 2 hours. The data is in the wave format with 16 kHz sampling rate and analog/digital conversion precision of 16 bits. The language model in this experiment is a bi-gram model which is trained by using all of transcriptions in the training data.

5.2. Feature extraction

5.2.1. Acoustic feature

In this experiment, we apply two kinds of acoustic features which are PLP and MFCC. They are extracted by HTK [16] tool using filter-bank of 300Hz-9000Hz, frame period of 10ms, and analysis window length of 25ms. Each feature vector contains 42 dimensions of 13 coefficients MFCC/PLP plus 1 energy, the first and second derivatives:

PLP/MFCC={13PLP/13MFCC + Energy, $\Delta, \Delta\Delta$}.

5.2.2. Tonal feature

As we described in section 4, an original pitch feature is extracted by AMDF or NCC method using Snack tool [18] with low-pass filter bank of 60Hz-380Hz, the analysis window length of 25ms, and frame period of 10ms. Using the approach of ToneF1, all the values in the unvoiced regions are replaced by zero values. By this way we acquired two

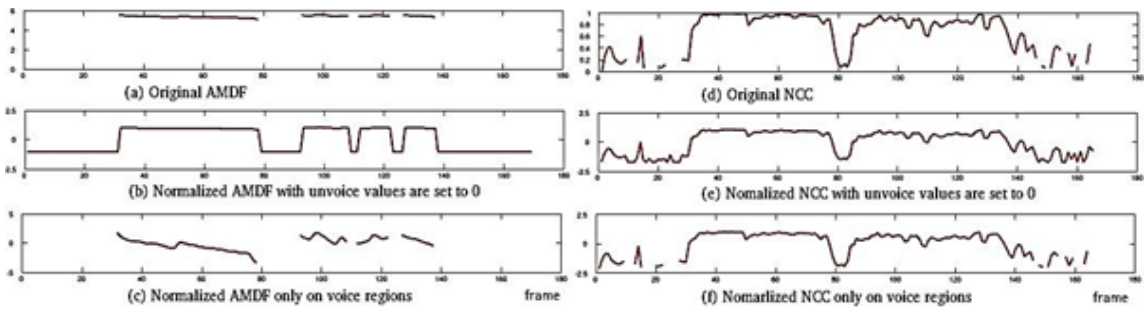

Figure 4. Output pitch contours of a utterance sample "chắc chữa được bách bệnh"

pitch features which are assumed as the continuous features called AMDF and NCC. Two other pitch features called AMDF_MSD and NCC_MSD acquired by using the approach of ToneF2. AMDF_MSD and NCC_MSD contain both special symbol for unvoiced regions and real values for voice regions. These features are modeled by MSD-HMM for comparing to AMDF and NCC which are modeled by normal HMM without MSD. Each pitch feature vector of all kinds consists of three dimensions of pitch, first and second deviation: AMDF/AMDF_MSD/NCC/NCC_MSD=$\{F0, \Delta F0, \Delta\Delta F0\}$.

5.2.3. Acoustic model

The acoustic models of tied states are trained by using HTS [19] tool. An MSD-HMM prototype described in Figure 1 has five states included the beginning and ending states. The input feature for this model has four independent streams which are acoustic feature, $F0, \Delta F0$, and $\Delta\Delta F0$ respectively. While the first stream containing the acoustic feature is one of PLP or MFCC or a combination of PLP/MFCC with AMDF/NCC feature. This stream is molded without MSD using 16 Gaussian mixtures. The second, third and fourth streams containing AMDF_MSD or NCC_MSD feature which are modeled by MSD using 2 Gaussian mixtures for each stream. In addition to apply MSD-HMM, we also applied an HMM prototype without MSD, where all the parameters for training this prototype are the same expect the deference that is the input feature contains only one stream instead of four streams.

5.3. Experiments

5.3.1. Task 1: Experiment on tonal phoneme

For evaluating the performance of the tonal phoneme (PS2) and non-tonal phoneme (PS1), we train four baseline systems without MSD based on these phoneme sets by using input feature which is PLP or MFCC. The system using PS1 has 2179 tied-states phonemes models and the system using PS2 has 6609 tied-states phonemes models. All of the systems are trained using 16 Gaussian mixtures. The experiment results are presented on word accuracy (ACC) which are shown in Table 4 from ID 1 to ID 4. The results show that the tonal phoneme set is better than non-tonal phoneme set. PS2-based systems improved 0.61% ACC on MFCC feature and 0.81% ACC on PLC feature compared to the systems using PS1.

Table 4. Summary of results obtained on tonal phoneme and MSD experiments

ID	System	Feature kind	Input feature	Phoneme set	ACC(%)
1	Baseline	Acoustic	MFCC	PS1	77.70
2			PLP		76.77
3			**MFCC**	PS2	**78.31**
4			PLP		77.58
5	Without SMD	Acoustic+ Tonal feature (ToneF1)	**MFCC+NCC**	PS2	**80.26**
6			MFCC+AMDF		76.10
7			PLP+NCC		79.09
8			PLP+AMDF		74.34
9	With MSD	Acoustic+ Tonal feature (ToneF1 and ToneF2)	MFCC+NCC_MSD	PS2	77.64
10			MFCC+AMDF_MSD		80.37
11			PLP+NCC_MSD		76.47
12			PLP+AMDF_MSD		79.78
13			**MFCC+NCC+AMDF_MSD**		**80.80**
14			PLP+NCC+AMDF_MSD		79.71

5.3.2. Task 2: Experiment on continuous using ToneF1

In these experiments, we want to evaluate the effect of AMDF and NCC on the systems

which are applied a normal HMM without MSD. The results will be compared to MSD-based systems. We apply the same HMM prototype as in the previous experiments on Task 1, but with the difference in the input feature. The input feature is a combination of acoustic feature and continuous tonal feature (ToneF1). Four systems are trained by combining respectively MFCC, PLP with AMDF, NCC to obtain four different input features. The results are shown in the Table 4 from ID 5 to ID 8. We obtain the best ACC on the system using feature combination of MFCC and NCC (ID 5 in Table 4) which improved 1.95% ACC compared to the best baseline system.

The systems using the combination feature of acoustic feature and NCC (MFCC/PLP+NCC) give the results better than the systems using combinations with AMDF (MFCC/PLP+AMDF). The combinations with AMDF even made the performance worse than the baseline systems. For example, the MFCC baseline system (ID 3 in Table 4) has 78.31% ACC, while the MFCC+AMDF system (ID 6 in Table 4) has 76.10% ACC. These results show that it is necessary to apply some methods for interpolating artificial values into the unvoiced regions of pitch feature in case of modeling by the normal HMM, otherwise the pitch feature would not improve the quality of recognition system.

5.3.3. Task 3: Experiment on MSD-HMM model using ToneF2

In MSD experiments, we apply an MSD prototype as described in section 2 using input feature with four independent streams. The first stream is the acoustic feature which is modeled by a normal HMM using 16 Gaussian mixtures. The second, third and fourth streams are tonal feature ($F0$), $\Delta F0$ and $\Delta\Delta F0$ which are modeled by MSD using 2 Gaussian mixtures for each stream. The tonal features have been applied in this task are not similar to the tonal features in Task 2. In these experiments, the tonal features consist of both continuous and discrete values, and modeled by the MSD.

There are six systems to be trained, the first four systems applied features which are the combinations of acoustic features (MFCC or PLC) with ToneF2 features (AMDF_MSD or NCC_MSD) respectively (MFCC/PLP+AMDF_MSD, MFCC/PLP+NCC_MSD, from ID 9 to ID 12 in Table 4), and two other systems are similar except the difference in the first stream which is a combination of MFCC or PLP with NCC feature (MFCC+NCC+AMDF_MSD, PLP+NCC+AMDF_MSD, ID 13 and ID 14 in Table 4). The results are shown in Table 4. We obtain the best number on the system using combination feature of MFCC, NCC and AMDF_MSD (ID 13 in Table 4) which improved 2.49% ACC compared to the best baseline system, and improved 0.54% ACC compared to the best system applying the normal HMM.

In the Task 2, the combination features with NCC give improvement by using a normal HMM, but in these experiments by applying MSD, contrariwise, AMDF_MSD give improvement. All of systems using the combination features with AMDF_MSD give better results than the combination features, which are combined with NCC_MSD. In fact, the combination features with NCC_MSD made the recognizing quality worse than the baseline systems. For example, the system ID 9 in Table 4 using MFCC+NCC_MSD feature give 77.64% ACC, while the baseline system ID 3 in Table 4 using MFCC gave 78.31% ACC.

6. SUMMARY AND CONCLUSIONS

In this paper, we have proposed an approach of MSD for Vietnamese recognition. An MSD-HMM prototype with four independent streams of input feature are applied for modeling the tonal phonemes. We also have conducted an approach for normalization. It is applied to tonal

feature which consists of both continuous and discrete values. The results show that MSD is effective on Vietnamese. It has improved ACC by 2.49% compared to the best baseline system, and 0.54% compared to the best system without MSD. The tonal phoneme set has improved ACC about 1% compared to non-tonal phoneme set, as shown in Table 4.

According to our observing of the experiments in Task 3. We have found that AMDF-based tonal feature is compatible with MSD, but NCC-based tonal feature is not. Since the NCC-based tonal feature contains less data samples for unvoiced regions, there are not much enough samples for estimating the parameters of unvoiced subspace.

In this work, the MFCC-based features give the better results than the PLP-based features. By combining tonal feature with acoustic feature, it has improved ACC in both cases of applying MSD or without. It shows that the tonal feature is an important factor for tonal language recognition. In next investigation, we continue the research on how to extract a better Vietnamese tonal feature for applying MSD-HMM model.

REFERENCES

[1] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong, John-Paul Hosom, Vietnamese large vocabulary continuous speech recognition, *Proc. INTERSPEECH*, Lisbon, September 2005 (1172–1175).

[2] Thang Tat Vu, Khanh Nguyen Tang, Son Hai Le, Mai Chi Luong, Vietnamese tone recognition based on multi-layer perceptron network, *Conference of Oriental Chapter of the International Coordinating Committee on Speech Database and Speech I/O System*, Kyoto, December 2008 (253–256).

[3] Phu Ngoc Le, Eliathamby Ambikairajah, Eric H.C. Choi, Improvement of Vietnamese tone classification using fm and $mfcc$ features, *Proc. Computing and Communication Technologies (RIVF 2009)*, Da Nang, Vietnam, 2009 (1–4).

[4] Ngoc Thang Vu, Schultz T., Vietnamese large vocabulary continuous speech recognition, *Proc. Automatic Speech Recognition & Understanding (ASRU)*, Merano, 2009 (333–338).

[5] Florian Metze, Zaid A. W. Sheikh, Alex Waibel, Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, Van Huy Nguyen, Models of tone for tonal and non-tonal languages, *Proc. Automatic Speech Recognition & Understanding (ASRU)*, Olomouc, Czech Republic, December 2013.

[6] Tokuda K., Takashi Masuko, Noboru Miyazaki, Takao Kobayashi, Hidden Markov models based on multi-space probability distribution for pitch pattern modeling, *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix USA, Mar. 1999 (229–232).

[7] Tokuda K., Takashi Masuko, Noboru Miyazaki, Takao Kobayashi, Multi-space probability distribution HMM, *The Institute of Electronics, Information and Communication Engineers (IEICE) Technical Report*, Japan, Vol. E85-D, 2002 (455–464).

[8] Yao Qian, Frank K. Soong, A Multi-Space Distribution (MSD) and two-stream tone modeling approach to Mandarin speech recognition, *Proc. Speech Communication*, Beijing China, 2009 (1169–1179).

[9] Doan Thien Thuat, *Ngu am tieng Viet (Vietnamese Acoustic)*, Vietnamese National Editions, Second edition, 2003.

[10] Nguyen Dinh Hoa, *Tieng Viet khong son phan*, John Benjamins Publishing Company, 1997 (ISBN-1-55619-733-0).

[11] Vu Thanh Phuong, "The acoustic and perceptual nature of tone in Vietnamese", Ph.D. thesis, Australia National university, Canberra, 1981.

[12] Hansjorg Mixdorff, Nguyen Hung Bach, Hiroya Fujisaki and Mai Chi Luong, Quantitative analysis and synthesis of syllabic tones in Vietnamese, *Proc. INTERSPEECH*, Geneva, 2003.

[13] M.S. Han, K.O Kim, Phonetic variation of Vietnamese tones in disyllabic utterances tones, *Journal of Phonetics* **2** (1974) 223–232.

[14] Dung Tien Nguyen, Mai Chi Luong, Bang Kim Vu, Hansjoerg Mixdorff , Huy Hoang Ngo, Fujisaki model based $f0$ contours in vietnamese tts, *Proc. International Conference on Spoken Language Processing (ICSLP)*, Korea, 2004.

[15] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, H. Manley, Average magnitude difference function pitch extractor, *Acoustics, Speech and Signal Processing, IEEE* **22** (1974) 352–362.

[16] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours", Ph.D. Thesis, Polytechnic Institute of Brooklyn, Michigan, 1986.

[17] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland, *The HTK Book (for HTK version 3.4)*, Cambridge University Engineering Department, 2006.

[18] The snack sound toolkit, http://www.speech.kth.se/snack/.

[19] Hmm-based speech synthesis system, http://hts.sp.nitech.ac.jp/?Home, 2011.

Received on February 36, 2019

Revised on October 46, 2019