NEURAL NETWORK-BASED TONAL FEATURE FOR VIETNAMESE SPEECH RECOGNITION USING MULTI SPACE DISTRIBUTION MODEL

Nguyen Van Huy*
College of Technology - TNU

SUMMARY

This paper presents a new approach of integrating Bottleneck feature (BNF) which is used for extracting tone information, to adapt to Multi Space Distribution Hidden Markov Model (MSD-HMM) for Vietnamese Automatic Speech recognition (Vietnamese ASR). In order to improve the performance of tonal feature, the first point that we present is a progress for extracting tonal feature based on a bottle neck Multilayer Perceptron (MLP) network that so called tonal bottle neck feature. The second major point in this paper is that we describe an approach for adapting the TBNF to MSD-HMM model. A new building system was trained with the appropriated topology for BNF size and MLP topology of hidden layers for tone recognition. Experiments on new building recognition system with TBNF integration are done to compare to 1/ a baseline system using MFCC feature and normal HMM prototype of five states, and 2/ a MSD-HMM system with widely used for extraction pitch feature such as Average Magnitude Difference Function (AMDF). Recognition accuracy on the testing set is 80.69%, it improved 2.38% compared to the baseline system and 0.32% compared to the best MSD-HMM system using the standard pitch feature AMDF.

Keywords: *Multi space distribution, bottle neck feature, tonal bottle neck feature, Vietnamese tone recognition, pitch feature*

INTRODUCTION

Tonal languages like Vietnamese, Mandarin and Cantonese generally use tones to represent phone level distinction, which are therefore essential to distinguish between words. Such tone information is generated by excursions in fundamental frequency, a feature that most recognition systems today discard as irrelevant for speech recognition. Vietnamese is a tonal monosyllable language in which each syllable has only one of six tones. Vietnamese ASR that integrated tone recognition for large vocabulary continuous speech is only at the beginning phase of development. Recently, there were several results that proposed some approaches for tone recognition of Vietnamese 0-0 but these approach model tones by applying a continuous tonal feature which can be obtained through the fundamental frequency F0, 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. The methods to extract tonal feature in the papers 0-0 try to fix errors in the unvoiced region or replace the unvoiced pattern by a random continuous value. In 0, another approach for Vietnamese ASR integrated tone recognition based on MSD-HMM by applying tonal phonemes was presented. With this approach, tonal phonemes using a combination of tonal and acoustic features were modeled, 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. In this paper, we describe how to integrate BNF which is used for extracting tone feature, to adapt to MSD-HMM for Vietnamese ASR. For this purpose, we present a process to 0 extract tonal feature based on a bottle neck MLP network that so called tonal bottle neck feature (TBNF). This TBNF can contain the variation information of F0 contour by concatenating more neighbor frames as input feature to MLP

* *Tel: 0968 852824, Email:huynguyen@tnut.edu.vn*

network. BNF is a kind of probability feature, it is computed based on a trained MLP. Based on careful experiments on training on size of BNF (see Table 4) and we found an appropriate BNF size, which is used afterward to define a topology of hidden layers sizes of trained tone recognition MLP network. Then for the paper's purpose, we integrate the TBNFs including both voiced/unvoiced information that has been trained on MLP topology described above to test MSD-HMM system in order to compare 1/ to only baseline HMM with MFCC, 2/ to MSD-HMM with widely used for extraction pitch feature such as Average Magnitude Difference Function (AMDF).

This paper is organized as follows: In Section 2, the basics of MSD and MSD-HMM are described. In Section 3, we present characteristics of Vietnamese tones. A proposed tonal bottle neck feature (TBNF) and its extraction process that is appropriated for the MSD-HMM model is presented in Section 4. The experimental results are given in Section 5. We conclude the paper in Section 6 with the summary and discussion of this study.

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, a 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 solve this problem. The first approach is to replace unvoiced patterns by heuristic values, and then model these patterns by using the continuous HMM. The second approach is to adapt 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 00 without any heuristic information and it was successfully applied for Mandarin 0. It can model the feature that consists of both continuous and discrete values, so we do not need to use any method for interpolation of artificial values into the unvoiced regions of pitch. The observation probability function of vector x in the normal HMM is defined by expression (1), then it is redefined by expression (2) in MSD-HMM model.

$$b_i(x) = \mathrm{N}_i(x) \quad (1) \qquad b_i(o) = \sum_{g \in I} \omega_{ig} \quad (2)$$

$$\mathrm{N}_{ig}(x),\ i=1,2,..,N$$

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

BASIC OF VIETNAMESE TONES

Vietnamese is a tonal monosyllable language, each syllable may be considered as a combination of Initial, Final and Tone components. 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 0. 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. Syllables with a closure coda can only go with rising tones and drop tones 0 which ending with stop consonants have F0 contours similar to rising and falling tones of other syllables, but they rise or drop more sharply 00. Therefore, most linguists who study Vietnamese acoustics claim that the Vietnamese language contains 8 different tones base on F0 contours. In this paper we only experiment on six lexical tones.

TONAL BOTTLE NECK FEATURE

Bottle neck feature

Bottle neck feature (BNF) is a kind of probability feature. It is computed based on a trained MLP network which usually has five layers where the size of the middle layer (or called as bottleneck layer) is small. For computing BNF, only three first layers are used. The size of BNF doesn't depend on the size of input feature, and it can be used to model by HMM directly. Many researches show that BNF really helps to improve performance of ASR systems. BNF is usually better than a normal acoustic feature, since it is classified by a MLP network. It also contains the time context of input feature, since the input feature for MLP network can be a combination of many neighbors of a vector. To maintain these advantages, we are going to apply MLP for extracting tonal feature, and adapt it to the MSD model.

Tonal bottle neck feature

Similar to other tone languages, Vietnamese tones can be represented by F0 contour. But the time span of F0 contour is usually longer than the time span of an F0 frame, therefore each individual F0 frame does not contain the variation information of F0 contour within the duration of a syllable. In order to achieve a better tonal feature we present a progress to extract a tonal bottle neck feature (TBNF), so-called, based on a bottle neck MLP network. This TBNF would contain more variation information of F0 contour?, since the input feature used for calculating TBNF is a concatenation of neighbor frames. TBNF then is adapted to the/an MSD model. A trained bottle neck MLP network of five layers is used to extract TBNF. The first layer is the input layer, the last layer is the output layer for Vietnamese tone targets, the middle layer is the bottle neck (BN) layer, and other layers are hidden layers. At first, the input feature is forwarded from the first layer to the BN layer for calculating a raw TBNF. This raw TBNF, activation values at BN layer, only consists of continuous values, and is considered as a kind of feature that its values belongs to only one space (so called voice space), whereas the MSD model is applied to features consisting of more than one space. To make TBNF suitable for MSD model, for this work, another space (so called unvoiced space) presenting for no-tone values in silence or unvoiced regions was added. The probability values of no-tone targets at the output layer are used to decide which TBNF frames belongs to voice or unvoiced space. If this probability is the maximum, then the TBNF will be set to no-tone (NT) label, otherwise keep the same the TBNF values. The raw TBNF, after applying voice/unvoiced decision, was used as TBNF in this work. Figure 3 and expression (3) present for this approach.

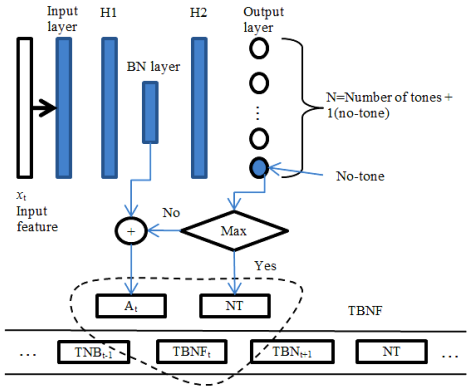

Figure 3: *Extracting tonal bottle neck feature included voice/unvoiced decision*

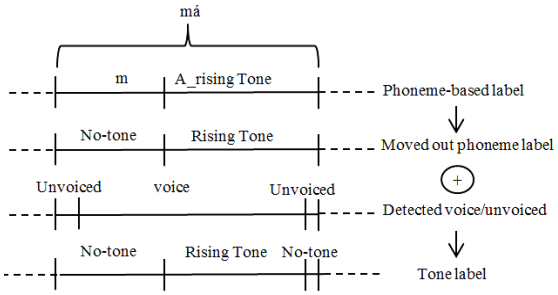

Figure 4: *An example of tone label assigning*

$$TBNF_t(x_t)=\begin{cases} NT, \text{ if } \arg ument(\max\{P_{t0},P_{t1},...,P_{tN}\})=I_{NT} \\ A_t, \text{ otherwise} \end{cases} \quad (3)$$

where x_t is input feature at time t, A_t is activation value of x_t at the bottle neck layer after forwarding from the first layer. P_{tj} ($j=1..N$, N is number of classification targets at output layer) is probability of x_t belong to target j. I_{NT} is index of no-tone target.

EXPERIMENTS SETUP AND RESULTS

For all of the experiments reported in this paper, we apply the tonal phoneme set which proposed in 0. Every Nucleus phoneme and Coda phoneme in the Final part of each syllable are combined with a tonal symbol according to its syllable. There are 152 tonal phonemes in this phone set. The language model used is a bi-gram model which is trained by using all of transcriptions in the training data.

Speech corpus

The data used in our experiments is the Voice of Vietnam (VOV) data which is a collection of story reading, mailbag, news 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. The VOV corpus is separated into training set of 17 hours and testing set of 2 hours. The data is in the wave format with 16 kHz sampling rate and analog/digital conversion precision of 16 bits.

Spectral feature

We apply two kinds of spectral features which are Mel Frequency Spectral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). They are extracted by HTK 0 toolkit using filter-bank of 300Hz-9000Hz, frame shift of 10ms, and analysis window length of 25ms. Each feature vector contains 42 dimensions of 13 coefficients for MFCC/PLP, 1 coefficient for energy, the first and second derivatives.

Pitch feature

The results from systems in 0 shown that pitch feature is extracted by Average Magnitude Difference Function (AMDF) 0 is suitable for MSD model, since AMDF contains enough samples for training parameters of unvoiced space. We maintain this approach to extract pitch feature and use it as a standard pitch feature with MSD model in order to compare its results to systems using TBNF. Pitch feature extracted by Normalized Cross-Correlation (NCC) 0 in systems 0, improved the accuracy when it was combined with MFCC or PLP on the systems without MSD. Therefore we decided to use NCC for extracting TBNF. Both AMDF and NCC (included its first and second derivatives) were computed by Snack toolkit 0 with low-pass filter bank of 60Hz-380Hz, the analysis window length of 25ms, and frame period of 10ms.

Baseline system

A system using MFCC feature and normal HMM prototype (without MSD, TBNF) of five states was used as a baseline system. This system was trained by HTK toolkit. There are 6609 tied-states tonal phoneme models with 16 Gaussian mixtures for each state. The result reported on percentage of accuracy (ACC). We got a baseline ACC of 78.31% on testing set (as shown in Table 3.)

MSD systems using standard pitch feature

An MSD-HMM prototype, described in 0 using input feature containing four independent streams, was applied for these experiments. The first stream, modeled by a normal HMM using 16 Gaussian mixtures, was spectral feature (MFCC/PLP). The second, third and fourth streams (F0, ΔF0, ΔΔF0) were AMDF feature modeled by MSD using 02 Gaussian mixtures. Two systems were trained by using HTS toolkit 0. The results are shown in Table 3. We obtained the best number on the system using combination of MFCC and AMDF that improved ACC of 2.06% compared to the baseline system.

Extracting TBNF

Tone label assigning

Firstly, the baseline system was used to realign all of the training data to get phoneme-based label. Then tone label was obtained by removing phoneme symbol except tone symbol. As the phoneme set was created in 0, because the tone symbol was supplemented in the Final part of each syllable, so tone would be considered that it exists in whole of the Final part. This is not in fact correct, because pitch representing tone does not exist in the unvoiced region. The Final part, in Vietnamese, could be decomposed into Onset, Nucleus and Coda whereas the Coda could be a consonant or a semi-vowel which could be unvoiced phonemes. Even whole of the Final part is a voice region, it could also be affected by the previous or next phoneme which could be unvoiced phoneme in a continuous utterance. To fix this problem, we detected voice and unvoiced regions for training data and rewrote tone label afterward. All of the frames in the Final part will be set to no-tone label, if they are in an unvoiced region. Figure 4 describes this progress for an example of syllable "má".

Training tone MLP network

For tone bottle neck MLP training, we applied an MLP topology with five layers. The BN layer is the middle layer. The size of input layer is 585 according to input feature of 13 neighbor vectors. Each vector is a normalized combination of MFCC with NCC (14MFCC + 1NCC, the first and second derivatives). Normalization was based on mean and standard deviation per utterance as expression (4). The size of output layer is 7 for classifying six tones and one no-tone targets. Firstly, we started choosing sizes of the first hidden layer (H1) and the third hidden layer (H2) are 1000 and 500 respectively, then trained MLPs with different sizes of BN. We wanted to find out the best size of BN layer for optimizing the performance of TBNF. Each trained MLP was used to compute TBNF afterward. It then combined with MFCC to train a simple system using normal HMM model. We decoded on the testing set to evaluate the performance. The results in Table 4 shown that the BN' size of 3 gives the best ACC. We continuously trained MLPs by keeping the same size of BN layer and changing sizes of H1 and H2. The results from the investigation of researches [18] [19] show that an MLP network giving the best cross-validation (CV) will give the best accuracy. Therefore in these experiments we try to find out the sizes of H1 and H2 which give the best CV. The best CV was 71.34% when the sizes of H1 and H2 are 1000 and 50 respectively. Results are shown in Table 5. Each line in "Topology" column in the table decibels sizes of MLP's layers respectively.

$$f_t = \frac{\log(x_t) - mean(X)}{Dev(X)}, \quad \left(Mean(X) = \frac{1}{T}\sum_{m=0}^{T-1} x_t, \; Dev(X) = \sqrt{\frac{1}{T}\sum_{T=0}^{T-1}(x_t - Mean(X))^2} \right) \quad (4)$$

where x_t is input feature, $X=\{x_0,..,x_t,..,x_T\}$, $t=0,..,T$ with T is length of an utterance.

Extracting TBNF

The MLP network, having CV of 71.34%, was used to compute the raw TBNF. Then it was normalized based on expression (4) and applied voice/unvoiced decision as described in section 4.2 to get TBNF.

Table 4: *Result of experiments on size of bottle neck layer*

BN size	ACC(%)
15	70.13
9	70.68
7	73.15
5	75.73
4	75.75
3	**76.53**
2	76.34

Table 5: *Experiment results on the sizes of hidden layers*

Topology	CV(%)
585-2000-3-500-7	70.09
585-1000-3-500-7	70.53
585-2000--3-100-7	70.72
585-2000-3-50-7	71.13
585-1500-3-50-7	71.22
585-1000-3-50-7	**71.34**
585-800-3-50-7	71.28

Table 6: *Summary of experiment results*

System	Input feature	ACC
Baseline	MFCC	78.31
Pitch feature	MFCC+AMDF	80.37 (+2.06)
	PLP+AMDF	79.78
TBNF	MFCC+TBNF	**80.69 (+2.38)**

MSD system using TBNF

A system was trained using almost the same parameters as MSD systems using standard pitch feature. There is only one difference that we used an MSD-HMM topology which has only two streams instead of four streams. The first stream is spectral feature MFCC using 16 Gaussian mixtures for each state. The second stream is TBNF using 4 Gaussian mixtures for each state. There are 6609 tied-states in this system. The ACC result on the testing set is 80.69% (as shown in Table 6). It improved by 2.38% compared to the baseline system and 0.32% compared to the best MSD system using the standard pitch feature.

DISCUSSION AND CONCLUSION

For the purpose of how to adapt BNF which is used for extracting tone feature, to MSD-HMM for Vietnamese ASR, we presented a process of extracting tonal feature based on a bottle neck MLP network that so called tonal bottle neck feature (TBNF) which included both voiced/unvoiced decision. In Table 1, based on carefully experiments on size of bottle neck layer, we show an appropriate size of BNF which is used afterward to define a topology of hidden layers size of trained tone recognition' MLP network. The experiments have shown that the first hidden layer and the third hidden layer is 1000 and 50 respectively give the best performance in term of cross validation (CV) accuracy. For the last experiment, we adapted TBNFs that were trained on MLP topology described above to test MSD-HMM system in order to compare to 1/ only baseline HMM with MFCC, 2/ to MSD-HMM with widely used for extraction pitch feature such as Average Magnitude Difference Function (AMDF). Experiment results show that on the testing set, accuracy is improved by 2.38% (80.69) compared to the baseline system (78.31) and 0.32% compared to the best MSD system using the standard pitch AMDF feature (80.37). In the next research, we will continue an investigation on how to extract a better Vietnamese tonal feature, namely integration several techniques for training acoustic and TBNF features, which could be classified by Linear Discriminant Analysis (LDA) before applying HMM-GMM.

REFERENCES

1. Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong, John-Paul Hosom, 2005, "Vietnamese large vocabulary continuous speech recognition", *Proc. INTERSPEECH*, Lisbon, pp. 1172-1175.

2. Thang Tat Vu, Khanh Nguyen Tang, Son Hai Le, Mai Chi Luong, 2008, "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, pp.253-256.

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

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

5. Nguyen Van Huy, Luong Chi Mai, Vu Tat Thang, Do Quoc Truong, 2014, "Vietnamese recognition using tonal phoneme based on multi space distribution", *Journal of Computer Science and Cybernetics*, Vietnam academy of science and technology, ISSN 1813-9663, pp. 28-38.

6. Tokudah K., Takashi Masuko, Noboru Miyazaki, Takao Kobayashi, 1999, "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, pp. 229-232.

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

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

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

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

11. M.S. Han, K.O Kim, 1974, "Phonetic variation of Vietnamese tones in disyllabic utterances tones", *Journal of Phonetics*, Vol. 2, pp. 223-232.

12. Dung Tien Nguyen, Mai Chi Luong, Bang Kim Vu, Hansjoerg Mixdorff , Huy Hoang Ngo, 2004, "Fujisaki model based f0 contours in vietnamese tts", *Proc. International Conference on Spoken Language Processing (ICSLP)*, pp.1429-1432, Korea.

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

14. M. Ross, H. Shaffer, A. Cohen, R. Freudberg, H. Manley, 1974, "Average magnitude difference function pitch extractor", *Acoustics, Speech and Signal Processing, IEEE*, Vol. 22, pp. 352-362.

15. B. S. Atal, 1986, *Automatic Speaker Recognition Based on Pitch Contours*, Ph.D. Thesis, Polytechnic Institute of Brooklyn, Michigan.

16. "Snack sound toolkit", http://www.speech.kth.se/snack/

17. 2011, "HMM-based speech synthesis system," http://hts.sp.nitech.ac.jp/?Home

18. Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, Van Huy Nguyen, Florian Metze, Zaid A. W. Sheikh, Alex Waibel, 2013, "Models of tone for tonal and non-tonal languages", *Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, pp. 261-266.

19. Kevin Kilgour, Christian Mohr, Michael Heck, Quoc Bao Nguyen, Van Huy Nguyen, Evgeniy Shin, Igor Tseyzer, Jonas Gehring, Markus Muller, Matthias Sperber, Sebastian Stuker and Alex Waibel, 2013, "The 2013 KIT IWSLT Speech-to-Text Systems for German and English", *International Workshop on Spoken Language Translation (IWSLT)*, Germany.

TÓM TẮT

ĐẶC TRƯNG THANH ĐIỆU DỰA TRÊN MẠNG NƠRON TRONG NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT SỬ DỤNG MÔ HÌNH PHÂN BỐ ĐA KHÔNG GIAN

Nguyễn Văn Huy[*]
Trường Đại học Kỹ thuật Công nghiệp – ĐH Thái Nguyên

Bài báo trình bày một cách tiếp cận mới về việc cải tiến đặc trưng Bottleneck và sử dụng nó cho mô hình Markov ẩn với hàm phân phát tán đa không gian (HMM-MSD). Để nâng cao chất lượng đặc trưng thanh điệu trong nhận dạng tiếng nói tiếng Việt bài báo trình bày quy trình sử dụng mạng nơron đa lớp có cấu trúc cổ trai để trích chọn đặc trưng. Sau đó nghiên cứu đề xuất một phương pháp mới để cải tiến đặc trưng này cho nó tương thích với mô hình HMM-MSD. Kết quả thử nghiệm trên đặc trưng mới được so sánh với hai hệ thống. Một là hệ thống cơ sở sử dụng đặc trưng ngữ âm và mô hình Markov ẩn thông thường. Hệ thống thứ hai sử dụng mô hình HMM-MSD và đặc trưng thanh điệu thông thường. Việc so sánh với hai hệ thống này nhằm chỉ ra hiệu quả của đặc trưng được tính toán theo phương pháp mới. Các kết quả thí nghiệm cho thấy đặc trưng mới trên mô hình HMM-MSD đã cho kết quả nhận dạng tốt hơn hệ thống cơ sở 2.38%, và tốt hơn hệ thống HMM-MSD thông thường là 0.32%.

Từ khóa: *Hàm phân bố đa không gian, đặc trưng Bottleneck, đặc trưng thanh điệu, nhận dạng thanh điệu tiếng Việt*

Ngày nhận bài:20/6/2015; Ngày phản biện:06/7/2015; Ngày duyệt đăng: 30/7/2015
Phản biện khoa học: *PGS.TS Nguyễn Duy Cương - Trường Đại học Kỹ thuật Công nghiệp - ĐHTN*

[*] *Tel: 0968 852824, Email:huynguyen@tnut.edu.vn*