社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ICS－138　(44)
2004／12／7

HMM/ANN System for Vietnamese Continuous Digit Recognition

Dang Ngoc Duc,[†] John-Paul Hosom,[††] Luong Chi Mai[†††]
and Vu Tat Thang[†††]

The study of a system for Vietnamese continuous digit recognition is described. The CSLU Toolkit was used to develop and implement hybrid HMM/ANN recognition systems. Experiments were done with a corpus of 442 sentences with 2340 words, which were extracted from two telephone-speech corpora: "22 Language v1.2" and "Multi-Language Telephone Speech v1.2". In our experiments, a context-dependent phoneme recognizer has achieved better recognition performance than a context-dependent demi-syllable recognizer and a context-independent phoneme recognizer. Among feature sets applied to the context-dependent phoneme recognizer, the set of 12 PLP features with CMS, energy and corresponding delta values has achieved the best recognition result (96.83% word accuracy and 87.67% sentence correct).

1. Introduction

Automatic speech recognition (ASR) is one branch of the field of artificial intelli-gence, dealing with automatic knowledge acquisition, models of human language, (acoustic) pattern recognition, and adaptation. Despite decades of research in ASR, acceptable "real-world" performance on even simple tasks remains, in some cases, elusive. In this paper, we describe research on the recognition of Vietnamese digits (the numbers zero through nine) spoken over the telephone channel without deliberate pauses between each word. Despite the small vocabulary size, this task is considered challenging because (a) the telephone channel has a severe impact on ASR perform-ance and (b) statistical models of the frequency of word combinations that improve performance on other tasks can not be applied to digit recognition. The task is then necessarily focused on pattern recognition in a noisy environment.

The purpose of the work reported here is to apply language-independent ASR tech-niques to the recognition of Vietnamese digits, to investigate the effect of different types of phonetic units on recognition performance, and to evaluate feature sets used in classification. High accuracy on the digits task will enable better automatic spoken-word acquisition, and might provide new techniques in pattern recognition that can be applied to fields other than speech recognition. (For example, techniques used in handwriting recognition are very similar to those in speech recognition.)

The CSLU Toolkit [4] was used in this work to carry out two experiments. In the first experiment, we compared the recognition performance of three recognizers based on three different basic speech units: context-dependent demi-syllables, context-independent phonemes and context-dependent phonemes. In the second experiment, we studied the effects of different feature sets to find the most suitable feature set for Vietnamese digit recognition. Eight feature sets were applied to context-dependent phoneme recognizers.

2. Basic Phonetic Structure of Vietnamese

Vietnamese is a monosyllable tonal language. Each Vietnamese syllable may be con-sidered a combination of Initial, Final and Tone components. The Initial component is always a consonant, or it may be omitted in some syllables. There are 21 Initials in Vietnamese. There are 155 Final components in Vietnamese[1] and the Final may 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 is 1 Onset, 16 Nuclei and 8 Codas in Vietnamese.

[†] Alcatel NSV, Vietnam Post and Telecommunication.
[††] Center for Spoken Language Understanding (CSLU-OGI).
[†††] Institute of Information Technology, Vietnamese Academy of Science and Technology.

The Tone is a super-segment and contains all parts of a syllable. There are six dis-tinct tones in Vietnamese, and they can affect word meaning; six different tones ap-plied to a syllable can result in six distinct words. The Initial, Tone, Onset, Nucleus and Coda may be combined together to make a syllable; however not all combinations are possible. There are a total of 18958 pro-nounceable distinct syllables in Vietnamese [1].

3. Corpus

The corpus used in this work for training, developing and testing our recognizers, consists of 442 sentences with 2340 words. It was extracted from two corpora from the Center for Spoken Language Understanding (CSLU): "22 Language v1.2", and "Multi-Language Telephone Speech v1.2".

Each sentence of the corpus consists of a number of digits in Vietnamese from 0 to 9. The sentences were recorded from 208 speakers (78 females and 130 males), who recited their telephone numbers, street addresses, ZIP codes or other numeric informa-tion over the telephone network in a natural speaking manner. The data were collected from different environments and may contain a noticeable amount of noise and other "real-life" aspects such as breath, glottalization, and music. The corpus was digitized at an 8000 Hz sampling rate with A/D conversion precision of 8 bits. All the sentences in the corpus have been time-aligned and transcribed at the phonetic level.

4. Experiment

4.1 Recognition System

The recognizers in this work were trained, developed and tested by the use of the CSLU Speech Toolkit [5], which is freely downloaded for research purposes from the OGI Web site. The hybrid HMM/ANN architecture, in which the phonetic likelihoods are estimated using a neural network, was chosen for all of our experiments. The document [4] was used as a guide for carrying our experiments.

Three-fifths of the data was randomly chosen for the training set, another one-fifth of the data was used for the development set, and the other one-fifth was used for the test set. The same data for training, developing, and testing was used for all experi-ments. The feature vectors were computed from the hand-labeled training data for each 10ms frame. The feature set contains features of the frame to be classified and features of frames at -60, -30, 30 and 60ms relative to the given frame.

The feature vectors were used for training a three-layer feed-forward neural net-work with an error back-propagation procedure. The neural network has 200 hidden nodes. The number of input nodes depends on the number of features (130 or 195 nodes); the number of output nodes depends on the number of categories of each rec-ognizer. The training was adjusted by negative penalty modification as described in [8]. The training was done for 30 iterations.

In the context-independent recognizer, each category of output of the network is a basic phonemic speech unit. In the context-dependent recognizer, basic speech units are split into 1, 2, or 3 categories depending on the length of phoneme and its influ-ence from the surrounding context. Each category is trained for different preceding and following phonetic contexts. Some contexts may be grouped together to form broad categories. The development set was used for evaluating the trained networks to find the best iteration. The best neural network was used for "forward-backward" (FB) training to improve the recognition results.

In the "forward-backward" (FB) training, the training strategy proposed by Yan et al. [6] was applied. In this method, the targets used to train the neural network are derived from posterior state occupation probabilities. The forward-backward re-estimation algorithm was used to regenerate the targets for training sentences. The neural network trained with hand-labeled data was used for the initial neural network. Unlike other hybrid systems, this hybrid HMM/ANN used within-phone model transitions. The training was finished after doing "forward-backward" training two times and the best FB2 recognizer was found for testing.

4.2 Experiment 1

In this experiment, we compared the recognition performance of three systems, based on different basic speech units: context-dependent demi-syllables, context-independent phonemes and context-dependent phonemes.

Table 1 Acoustic units for demi-syllable and phoneme recognition systems.

English	Vietnamese	Initial-Final	Phoneme
zero	khoong	/kh//oong/	/kh//oo//ng/
one	mootj	/m//oot/	/m//oo//te/
two	hai	/h//ai/	/h//a//i/
three	ba	/b//a/	/b//a/
four	boons	/b//oon/	/b//oo//n/
five	nawm	/n//awm/	/n//aw/m/
six	saus	/s//au/	/s//a//u/
seven	baayr	/b//aai/	/b//aa//i/
eight	tams	/t//am/	/uc//t//a/m/
nine	chins	/ch//in/	/uc//ch//i/n/

The first recognizer was a context-dependent demi-syllable recognizer. The Initial was defined as right dependent to take into account co-articulation effects from the first vowel in Final on Initial. The Final was split into 3 categories. The list of Initials and Finals in this recognizer is described in Table 1.

The second recognizer in this experiment was a context-independent phoneme rec-ognizer. Each phoneme was defined as one part. A list of phonemes is provided in Table 1, in which the unvoiced closure /uc/ is inserted in front of the unvoiced stops.

The last recognizer used in this experiment was based on context-dependent pho-nemes. Each phoneme was divided into one, two or three parts. The vowels were split in three parts, and all stops are defined as one part because they are very short.

The same grammar was used for all three recognizers. This grammar allows any digit to follow any other digit with equal probability, and each digit may be separated by either silence or "garbage" [3]. All three recognizers for this experiment were trained, developed and tested using the same feature set: 12 MFCC coefficients with cepstral mean subtraction, plus energy and their delta (D) values.

4.3 Experiment 2

In this second experiment, we applied different feature sets to our context-dependent phoneme recognizer. The 13 PLP coefficients (PLPC13) and 13 MFCC coefficients (MFCC13) were computed with one of two preprocessing techniques: RASTA (RelAtive SpecTrAl) or CMS (Cepstral Mean Subtraction). The delta-delta (D2, or acceleration) values were also added to the feature set. The motivation of this experi-ment was to study the influence of feature extraction on recognition performance.

5. Results

Table 2 shows the results of Experiment 1 with word accuracy (WA) and sentence correct (SC) for the development set and test set. The context-dependent phoneme recognizer has better performance in comparison with the context-independent phoneme recognizer, demonstrating the effectiveness of context-dependent modeling. The context-dependent phoneme recognizer has better recognition accuracy than the con-text-dependent demi-syllable recognizer, showing that the basic phonetic unit suitable for continuous digit recognition is the context-dependent phoneme.

Table 2 Recognition performances of three recognizers: context-dependent demi-syllable, context-independent phoneme and context-dependent phoneme. " WA " indicates word-level accuracy (in percent), and " SC " indicates sentence-level correct (in percent).

basic speech unit	set	WA	SC
context-dependent demi-syllable	dev	93.80	79.22
	test	93.02	79.45
context-independent phoneme	dev	92.99	76.62
	test	90.48	73.97
context-dependent phoneme	dev	95.69	81.82
	test	96.19	87.67

Table 3 shows the results of Experiment 2. The context-dependent phoneme recog-nizer with PLP13 (CMS) plus delta values achieves the best result with 96.83% word accuracy and 87.67% sentence correct.

It should be noted that Table 3 also demonstrates that the addition of D2 in the feature set does not improve significantly the performances of the recognizers.

Table 3 Recognition performance of the context-dependent phoneme recognizer with eight different feature sets.

feature set	set	WA	SC
mfcc13(cms)+D	dev	95.69	81.82
	test	96.19	87.67
mfcc13(rasta)+D	dev	93.80	77.92
	test	92.38	72.60
plpc13(cms)+D	dev	96.23	84.42
	test	96.83	87.67
plpc13(rasta)+D	dev	92.99	75.32
	test	94.60	80.82
mfcc13(cms)+D+D2	dev	96.50	87.01
	test	95.87	86.30
mfcc13(rasta)+D+D2	dev	95.15	81.82
	test	93.02	75.34
plpc13(cms)+D+D2	dev	96.50	85.71
	test	96.83	87.67
plpc13(rasta)+D+D2	dev	92.99	74.03
	test	94.29	76.71

6. Conclusions

In this paper, we have presented our study on continuous digit recognition for Vietnamese over the telephone line. The results show that context-dependent demi-syllable recognition has lower accuracy in comparison with our context-dependent phoneme recognizer. The context-dependent phoneme recognizer has better recogni-tion results in comparison with the context-independent phoneme recognizer. Thus, context-dependent phonetic units are better suited to continuous Vietnamese digit recognition. Furthermore, this investigation of the type of sub-word units points to the need to carefully choose the basic units of classification.

We also found in our experiments that among the feature sets used in context-dependent phoneme recognizers, the feature set with 12 PLP coefficients with CMS, plus energy and their delta values achieves the best result. In comparison with con-tinuous digit recognition over the telephone for English (using a much larger number of speakers for training) [3], our system has comparable performance: 96.83% word accuracy and 87.67% sentence correct.

In future research, more experiments with larger amounts of data need to be con-ducted to further confirm our findings. Also, experiments are planned to include in-formation on pitch contours to further improve accuracy.

References

1) Vu K. B., Trieu T.T.H and Bui D.B: Am tiet tieng Viet kha nang hinh thanh va thuc te ung dung, *Proc. of conference in IT*, Institute of IT, (2001).
2) Jim J.W, Li D. and Jacky C: Modeling context-dependent phonetic units in a continuous speech recognition system for Mandarin Chinese, *Proceeding of ICSLP '96*, (1996).
3) Hosom, J.P., Cole, R.A, and Cosi, P.: Improvements in Neural-Network Training and Search Techniques for Continuous Digit Recognition, *Australian Journal of Intelligent Information Processing Systems (AJIIPS)*, vol. 5, no. 4 (Summer 1998), pp. 277-284.
4) Hosom, J. P., Cosi, P. Cole, R., Fanty, M., Schalkwyk, J., Yan, Y. and Wei, W.: Training Neural Networks for Speech Recognition, http://cslu.cse.ogi.edu/tutordemos/nnet training/tutorial.
5) http://cslu.cse.ogi.edu/toolkit.
6) Yan, Y., Fanty, M and Cole, R.: Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets, *In Proceedings ICSDDP97*, (1997).
7) Lander. T.: CSLU Labeling Guide, Center for Spoken Language Understanding, Oregon Graduate Institute, (1997).
8) Wei, W and Van Vuuren, S.: Improved Neural Network Training of Inter-Word Context Units for Connected Digit Recognition, *In Proceedings of International Conference on Acoustic Speech and Signal Processing (ICASSP '98)*, Seattle, Washington, May 1998, Vol. 1, pp. 497-500.