International Journal of Computational Linguistics and Applications, vol. 5, no. 2, 2014, pp. 153–164
Received 17/08/2014, accepted 4/11/2014, final 14/11/2014.
ISSN 0976-0962, http://ijcla.bahripublications.com

Evaluating Prosodic Characteristics for Vietnamese Airport Announcements

LAM-QUAN TRAN,[1] ANH-TUAN DINH,[2]
DANG-HUNG PHAN,[2] AND TAT-THANG VU[2]

[1] *Vietnam Aviation Institute, Vietnam*
[2] *Institute of Information Technology, Vietnam*

ABSTRACT

In most languages, the quality of a speech synthesis system relates directly to the diversity of language domain. Each domain, such as sports, entertainments, etc., has its specific grammar structures. The grammar structure plays as an important role for analyzing the prosodic information of utterances in each domain. In this research, we will analyze characteristics of prosodic information of airport domains in Vietnamese and detect most important characteristics related to the sentiment of Vietnamese Airport announcements.

1 INTRODUCTION

Advantages of the text-to-speech system have been utilized for many areas. To build a smooth voice, there are several statistical methods that have been widely researched. In statistical methods, Hidden Markov Model-based speech synthesis provides many benefits to build a high quality TTS system. With HMM, the TTS process is created by two main process: training data and synthesizing the input text achieved from users [3]. With HMM training, the voice can be created with small footprint of sound data [6], with lower than one hour of sound record. Moreover, based on the statistical method, HMM can model the co-articulation between consecutive sound units to provide smooth synthetic voice. However, one of the main disadvantages of HMM based speech synthesis is the naturalness. HMM training steps tend to neutralize the parameter of the synthesized output sound, including the F0, pitch and duration. The neutralization brings drawback that the voice is over-smooth and has low naturalness.

To overcome the low quality naturalness problem for HMM based TTS, the decision tree component has been improved for its task to detect phonemic and prosodic characteristics of the set of phonemes obtained from the input text [2]. Structure of the decision tree of HMM-based is language dependent, it depends on the grammar structure and prosodic information of input text. Due to that, the decision tree is also domain dependent. Its structure verified in each domains, such as sport, science, etc… However, in general Vietnamese Speech Synthesis System, the structure of decision tree lacks of prosodic information embedded in input text, brings the result that the voice output quality is still average in naturalness.

The main approach to produce a high quality decision tree for every language is to analyze characteristics of input text for a specific domain. In this research, we focus on analyzing information about the input text provided by a set of airport announcements [7]. Based on the analyzing, we detected characteristics provided by rules about prosodic, including part of speech, stress, and intonation of sample airport announcements. The result of this thesis is embedded to enhance the quality of the auto-announcement system of Vietnam Airline [7]. This paper includes three sections: prosodic analysis in airline announcements, improvement in HMM-based speech synthesis system and experiments.

2 Prosody Analysis in Airline Announcements

To understand the prosody phenomenon, the part of speech arrangements in the sentences are demonstrated. By observing F0 contour of training sentences, the relations between POS and Stress, POS and Intonation are established.

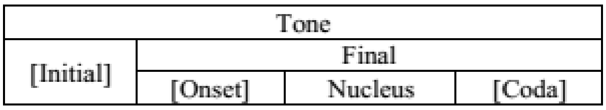
2.1 *Stress*

Stress is how a phoneme is underscored in a syllable. In languages such as Russian, English and French, stress is very important. However, in

Vietnamese and other tonal languages, the stress's role is less important than those of Russian, English… In Airline statement, a phoneme can be emphasized with long duration and loud voice. In training corpus, the syllable "*Nam*" in "*Việt Nam*", "*bay*" in "*chuyến bay*", "*đi*", "*số*" in "*quầy số*" or "*cửa số*" and "*VN*" is strengthen and has long durations. In addition, the noun such as space names and names of the flight are also underscored. The long duration appears when the announcers try to emphasize the important information followed the above words. The speakers are demanded to read the statement in a noisy and crowded environment like the airport and they have to ensure the important information can come to the customers.

According to Doan Thien Thuan [1], a syllable's structure can be demostrated as in Table 1. Vietnamese is a tonal monosyllable language, each syllable may be considered as a combination of Initial, Final and Tone components as 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 distinct pronounceable syllables in Vietnamese is 18958 [9] but the used syllables in practice are only around 7000 different syllables [1]. 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. By observation, the stress in the Airline speech utterances lies on Nucleus.

Table 1. Structure of Vietnamese syllable



Tone			
[Initial]	Final		
	[Onset]	Nucleus	[Coda]

2.2 *Intonation*

In our work, ToBI is used to transcribe the intonation in the training sentences. ToBI is a widely used transcription standards. It has been applied in prosodic analysis in many languages such as English, French, and Chinese… The primitive elements in ToBI are low (L) and high (H) tones. The melody of a sentence is divided into many elements. The elements are classified into two main groups: Phrase-final intonation and Pitch Accent.

Phrase-final intonation is the variation of spoken pitch at the end of a intonation phrase [2]. A sentence can have more than one intonation phrase. The L-L% (low-low) and L-H% (low-high) tags are used in transcription. L-L% is used at the end of a statement phrases and L-H% is used to mark the end of an emotional phrases and question utterances.

Pitch Accent is the failing and rising trend of pitch contour [2]. The failing trend is described by the H+L* (high-low) tag and L* (low) tag. L+H* (low-high) and H* (high) present the rising trend in the baseline of F0 contour. The following sentence in training set shows an example of using ToBI in Intonation Transcription. F0 contour of the sentence are partly shown in the Figure 1.

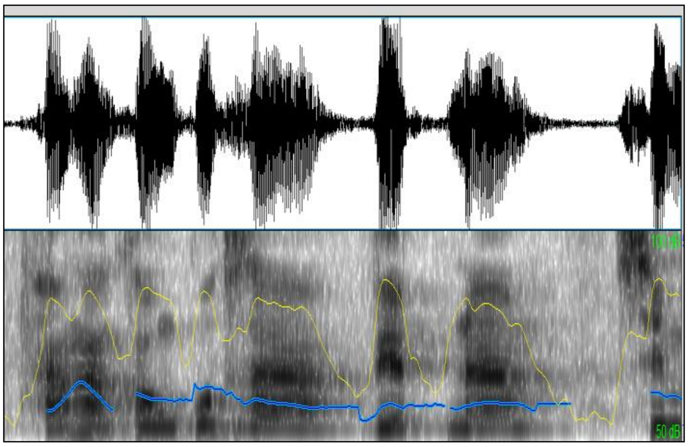
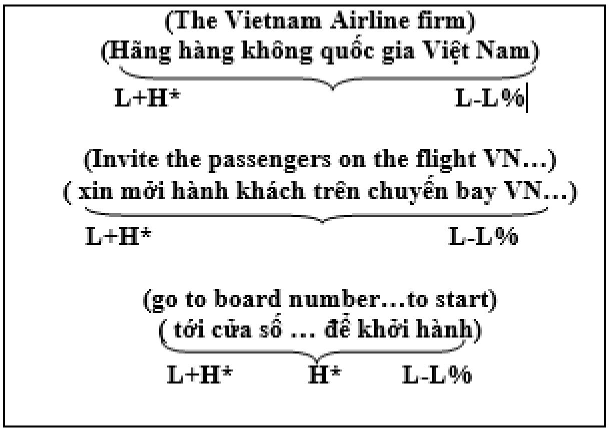


Fig. 1. Spectrogram and F0 contour of the sentence: "*Hãng hàng không quốc gia Việt Nam xin mời hành khách trên chuyến bay VN27 tới cửa số 08 để khởi hành.*" ("Vietnam Airline invites passengers on the flight VN27 please go to board 08 to start")

The relation between intonation transcription and F0 contour is very complicate. Some common relations are described in Table 2.

By observing the set of Airline announcement, the transition network of Vietnamese phrasal melodies can be established as in Figure 3.



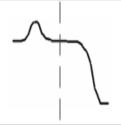
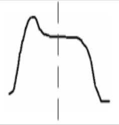
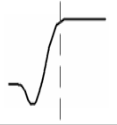
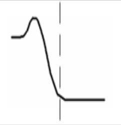
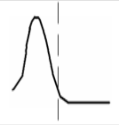
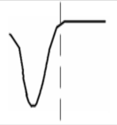
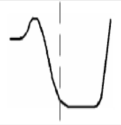
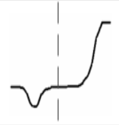
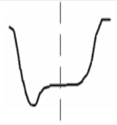
(The Vietnam Airline firm)
(Hãng hàng không quốc gia Việt Nam)
L+H* L-L%

(Invite the passengers on the flight VN…)
(xin mời hành khách trên chuyến bay VN…)
L+H* L-L%

(go to board number…to start)
(tới cửa số … để khởi hành)
L+H* H* L-L%

Fig. 2. Pitch accent analysis results of sample airline announcements

Table 2. Relation between intonation and F0 contour

F0 Contour	ToBI	F0 Contour	ToBI	F0 Contour	ToBI
	H* H-L%		L+H* H-L%		L* H-H%
	H* L-L%		L+H* L-L%		H+L* H-H%
	H* L-H%		L* L-H%		H+L* L-H%

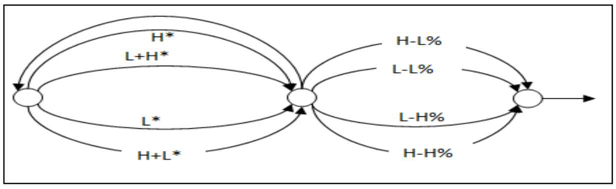


Fig. 3. ToBI grammar in Airline utterance

3 PROSODY IMPROVEMENT IN HMM-BASED SPEECH SYNTHESIS SYSTEM

3.1 *HMM-based Speech Synthesis System*

In HMM-based speech system, speech signals can be reconstructed from feature vectors. A feature vector consists of spectral parameters as Mel-Cepstral Coefficients (MCCs, or Mel-Frequency Cepstral Coefficients-MFCCs), duration, and excitation parameters such as fundamental frequency, F0. To understand the basic notation of HMM-based speech synthesis, we come to four concepts: Spectral modeling, Excitation modeling, State Duration modeling, Language-depent Contextual factors and Context-clustering decision tree.

In Spectral modeling, the MFCCs include energy component and the corresponding delta and delta-delta coefficients are used to represent the spectral. Sequences of Mel-frequency cepstral coefficient vector, which are obtained from speech database using a Mel-cepstral analysis technique, are modeled by continuous density HMMs. It enables the speech to be reconstructed from the coefficients by using the Mel Log Spectral Approximation (MLSA) filter. The MFCC coefficients are obtained through Mel-cepstral analysis by using 40-ms Hamming windows with 8-ms shifts. Output probabilities are multivariate Gaussian distribution [3].

In Excitation modeling, the excitation parameters are composed of logarithmic fundamental frequencies (logF0) and their corresponding delta and delta-delta coefficients. The continuous values in voice region and the discrete values in unvoice region are modeled by Multi-Space probability Distribution. [4]

State duration densities of phonemes are modeled by single Gaussian distributions [5]. Dimension of state duration densities is equal to the number of state of HMM, and the *n*-th dimension of state duration densities is corresponding to the nth state of phoneme HMMs. The duration of each state is determined by HMM-based speech synthesis system. State durations are modeled as multivariate Gaussian distribution [4].

In HMM-based speech synthesis approach, there are many Contextual factors (phone identity factors, stress-related factors, dialect factors, tone factors and intonation) that affect the spectral envelope, pitch and state duration. The only language-dependent requirements within the HTS framework are contextual labels and questions for context clustering. Some contextual information in Vietnamese language was considered as follows [3]:

Phoneme level:

– Preceding, current and succeeding phonemes.
– Relative position in current syllable (forward and backward)

Syllable level:

– Tone types of preceding, current and succeeding syllables.
– Number of phonemes in preceding, current and succeeding syllables.
– Position in current word (forward and backward).
– Stress-level.
– Distance to {previous and succeeding} stressed syllable.

Word level:

– Part-of-speech of {preceding, current, succeeding} words.
– Number of syllables in {preceding, current and succeeding} words.
– Position in current phrase
– Number of content words in current phrase {before, after} current word.
– Distance to {previous, succeeding} content words
– Interrogative flag for the word.

Phrase level:

– Number of {syllables, words} in {preceding, current, succeeding} phrases.
– Position of current phrase in utterance.

Utterance level:

– Number of {syllables, words, phrases} in the utterance

In many cases, a speech database doesn’t have enough contextual samples. In other word, a given contextual label doesn’t have its corresponding HMM in the training model set. Therefore, to solve this problem, a Context dependent clustering Decision Tree is applied to classify the phonemes. The question set can be easily extended to include more contextual information which helps the clustering becomes more detail. The questions are obtained from the phonetic and prosodic characteristics. In training phase, all speech samples of a phoneme with the same context are used to train a 5 state HMM for the context dependent phoneme. In synthesis phase, the decision tree is used to choose an appropriate HMM for each phoneme based on its context.

3.2 *HMM-based Speech Synthesis System*

To improve the naturalness of the synthetic airline statement, we will integrate more context-depend information specified for airline announcement. The adding information is about POS, stress and intonation of the airline utterances.

Figure 4 shows an example of the full context label of phoneme “tr” in the utterance “chúc ngủ ngon” (“good night” in English).

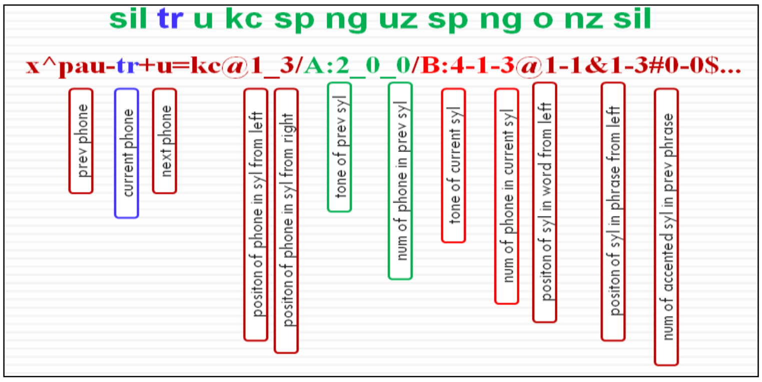
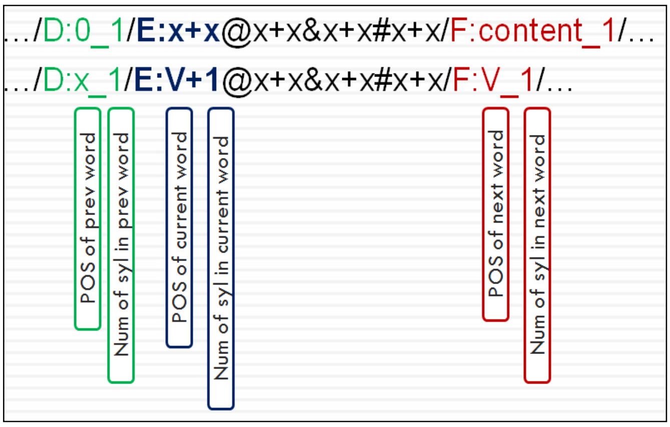

sil tr u kc sp ng uz sp ng o nz sil
x^pau-tr+u=kc@1_3/A:2_0_0/B:4-1-3@1-1&1-3#0-0$...
prev phone
current phone
next phone
positon of phone in syl from left
positon of phone in syl from right
tone of prev syl
num of phone in prev syl
tone of current syl
num of phone in current syl
positon of syl in word from left
positon of syl in phrase from left
num of accented syl in prev phrase

Fig. 4. Full context label of phoneme “*tr*”

Preceding, Current and Succeeding word’s POS information is added to the full context label. The information is obtained by using VietTagger and JvnTagger for POS tagging. Figure 5 shows the POS information in full context label.

In intonation transcription, the problem is to identify a melody phrase. The phrases are not always be sentences. Through observing the F0 contour of training speech, the common phrases of the utterances are:

- “*Hãng hàng không quốc gia Việt Nam*” (“The Vietnam Airline firm”)



```
.../D:0_1/E:x+x@x+x&x+x#x+x/F:content_1/...
.../D:x_1/E:V+1@x+x&x+x#x+x/F:V_1/...
```

Fig. 5. Added POS information in full context label

- “*xin mời hành khác trên chuyến bay VN*___” (“invite passengers on the flight VN__”)
- “*xin mời những hành khách cuối cùng đi*__” (“invite last passengers go to_”)
- “*trên chuyến bay*__” (“on the flight__”)
- “*tới cửa số*__” (“go to board number__”)
- “*trên chuyến bay VN*__” (“on the flight VN_”)
- “*khẩn trương tới cửa số*__ *để khởi hành*” (“hurry go to board number __ to start”)
- “*xin cảm ơn*” (“thanh you”)

Each of the phrases will have its Pitch Accent and Phrase Intonation Tone. The experiment shows positive result after Intonation transcription is added to full context label. The naturalness is improved.

Stress identification stays a problem in Vietnamese because stress does not play an important role in the language. In Section 2.2, some rules to identify the stress in Airline statement are shown. Based on the rules, the stressed phonemes and information about relative position and number of phonemes are added to the full context label.

4 Experimental Results

In order to evaluate the performance of new system after adding the new Prosodic information, an experiment was established. The experimental setup is shown in Table 4.

Table 4. Experimental setup.

Database	Northern Vietnamese female voice
Training/ test data	510 / 100 sentences
Sampling rate	16 kHz
Analysis window	25-ms width / 5–ms shift
Acoustic features	25 mel-cepstrum, log *F0*, delta and delta-delta
HMM topology	5-state, left-to-right, no skip HMM

4.1 *Subjective Test*

MOS test is used to measure the quality of synthesized speech signals, in comparison with natural ones. The rated levels are: bad (1), poor (2), fair (3), good (4) and excellent (5). In the test, 50 sentences were randomly selected. With 3 types of natural speech, synthetic speech without POS, stress and intonation and synthesized speech with POS, stress and intonation. The number of listeners are 50 people. The speech segments were played in random order in the test. Table 5 shows the MOS test results which were given by all the subjects. The MOS result implied that the quality of natural speech is excellent and the quality of synthetic voice with new prosodic information is better than the synthetic voice without the kind of information.

Table 5. Results of MOS test

Speech	Mean Opinion Score
Natural	5
Without POS, stress, intonation	3.26
With POS, stress, intonation	3.98

4.2 *Objective Test*

To evaluate the synthesis quality, Mel Cepstral Distortion (MCD) [8] is computed on held-out data set. The measure is defined as in Equation 1:

$$MCD=(10/ln10)\sqrt{2 * \sum_{i=1}^{25}(mc_i^t - mc_i^e)^2}, \qquad (1)$$

where mc_i^t and mc_i^edenote the target and the estimated mel-cepstral, respectively. MCD is calculated over all the MCEP coefficients, including the zeroth coefficient. Lesser the MCD value the better it is. Through observation, it's realized that a difference of 0.2 in MCD value produces difference in the perceptual difference in quality of synthetic speech. Table 6 shows the result of MCD evaluation.

Table 6. MCD results in comparison of 2 synthetic voices with natural voice.

Synthetic voice x Natural voice	MCD
Without POS, stress, intonation x Natural voice	6.47
With POS, stress, intonation x Natural voice	5.86

ACKNOWLEDGEMENTS This work was partially supported by ICT National Project KC.01.03/11-15 "Development of Vietnamese – English and English – Vietnamese Speech Translation on specific domain". Authors would like to thank all staff members of Department of Pattern Recognition and Knowledge Engineering, Institute of Information Technology (IOIT) - Vietnam Academy of Science and Technology (VAST) for their support to complete this work.

REFERENCES

1. Doan, T.T.: Vietnamese Acoustic, Vietnamese National Editions, Second edition (2003)
2. Phan, T.S., Dinh, A.T., Vu, T.T., Luong, C.M.: An Improvement of Prosodic Characteristics in Vietnamese Text to Speech System. In: Knowledge and Systems Engineering, Advances in Intelligent Systems and Computing, vol. 244, Springer, pp. 99–111 (2014)
3. Vu, T.T., Luong, M.C., Nakamura, S.: An HMM-based Vietnamese Speech Synthesis System. In: Proc. Oriental COCOSDA, Urumqi, China, pp. 108–113 (2009)
4. Tokuda, K., Masuko, T., Miyazaki, N, Kobayashi, T.: Multi-space Probability Distribution HMM. In: IEICE, vol. E85-D, no. 3, pp. 455–464 (2002)
5. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Duration Modeling in HMM-based Speech Synthesis System. In: Proc. ICSLP, vol. 2, Sydney, Australia, pp. 29–32 (1998)

6. Phung, T.N., Phan, T.S., Vu, T.T., Luong, M.C., Akagi, M.: Improving Naturalness of HMM-Based TTS Trained with Limited Data by Temporal Decomposition. In: IEICE Transactions on Information and Systems, vol. E96-D, no. 11, pp. 2417–2426 (2013)
7. Lam, Q.T., Dang, H.P., Anh, T.D.: Context-aware and Voice Interactive Search. In: Proc. 5th International Conference on Soft Computing and Pattern Recognition (2013)
8. Toda, T., Black, A.W., Tokuda, K.: Mapping from Articulatory Movements to Vocal Tract Spectrum with Gaussian Mixture Model for Articulatory Speech Synthesis. In: Proc. 5th ISCA Speech Synthesis Workshop, pp. 31–36 (2004)
9. Vu, K.B., Trieu, T.T.H., Bui, D.B.: Vietnamese Phonemic System: The Construction and Application. In Vietnamese. In: Proc. Celebration of 25 year establishing the Institute of Information Technology, Vietnam Academy of Science and Technology Conference, pp. 525–533 (2001)

Lam-Quan Tran
VNA: Vietnam Aviation Institute – Vietnam Airlines,
121 Nguyen Son, Long Bien, Hanoi, Vietnam
E-mail: <QuanTL.vai@vietnamairlines.com>

Anh-Tuan Dinh
Institute of Information Technology,
Vietnam Academy of Science and Technology,
18th Hoang Quoc Viet street, Hanoi, Vietnam
E-mail: <tuanad121@gmail.com>

Dang-Hung Phan
Institute of Information Technology,
Vietnam Academy of Science and Technology,
18th Hoang Quoc Viet street, Hanoi, Vietnam
E-mail: <pdhung3012@gmail.com>

Tat-Thang Vu
Institute of Information Technology,
Vietnam Academy of Science and Technology,
18th Hoang Quoc Viet street, Hanoi, Vietnam