HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation

NGUYEN Thi Thu Trang\textsuperscript{1,2}, Christophe D’ALESSANDRO\textsuperscript{1}, Albert RILLIARD\textsuperscript{1}, TRAN Do Dat\textsuperscript{2}
\textsuperscript{1} LIMSI-CNRS (UPR 3251), France
\textsuperscript{2} MICA, HUST - CNRS/UMI2954 - Grenoble INP, Vietnam
trangttt@soict.hut.edu.vn, cda@limsi.fr, albert.rilliard@limsi.fr, do-dat.tran@mica.edu.vn

Abstract
This paper presents the development and evaluation of an HMM-based TTS system for the modern Hanoi dialect of Northern Vietnamese, a tonal language. A study of specific phonetic and prosodic features of Hanoi Vietnamese is discussed. Consequences on the design of an HMM-based TTS system are derived. Using this knowledge, a TTS system, called VTed, is then developed under the Mary TTS platform. The second part of the paper is devoted to perceptual evaluations of Vietnamese speech synthesis. Three kinds of evaluations are considered necessary for quality assessment of this tonal language. The general MOS assessment, utterance-level intelligibility, and tone-level intelligibility tests are conducted on the VTed system under a “natural speech reference” condition. The results show 1.21 points difference between natural and synthetic speech for the MOS test, a 0.2% – 0.9% difference for the utterance-level intelligibility test, 23% on average and – depending on the tone type – from 0% to 37% difference for the tone-level intelligibility test. These results demonstrate the need for more specific works on tonal/prosodic level to improve automatic synthesis of Vietnamese and other tonal languages.

Index Terms: perceptual evaluation, text-to-speech, speech synthesis, tonal language, Vietnamese, Hanoi dialect

1 Introduction
Statistical parametric speech synthesis is most simply described as generating the average of some sets of similarly sounding speech segments. One instance of the techniques in this paradigm, the hidden Markov model (HMM) based speech synthesis, has recently been demonstrated to be very effective in synthesizing acceptable speech [1]. Although many works on HMM-based speech synthesis for tonal languages have been published [2] [3] [4], there is only few for Vietnamese. To the extent of our knowledge, there are two groups working on HMM-based Vietnamese speech synthesis: One from IoIT, Vietnam [5] [6] and another from Yunnan, China [7] [8]. They both adopted the HTS framework [9] and presented only the HTS core architecture, not the complete architecture of a Text-To-Speech (TTS) system.

This paper presents the entire design of an HMM-based TTS system for Vietnamese, and our study of phonetic and phonological features of this tonal language. VTed, a TTS system for Hanoi Vietnamese, is implemented following the proposed design. Both training and synthesis phases in VTed are performed automatically regarding tonal aspects.

There are a number of works on different methods and dimensions to evaluate TTS systems. A rigorous approach is directed towards the identification of auditory (perceptual) dimensions of speech quality, e.g. intelligibility, naturalness, loudness [10]. However, it seems that the evaluation of tone intelligibility has not received enough effort in preparations and thorough discussions. Evaluations in a TTS system for tonal languages are most often conducted to assess the naturalness or clearness using a MOS test, with or without a natural speech reference, such as in [2] [11] for Chinese, or in [12] [13] for Vietnamese. Some works have mentioned intelligibility but with only general results and without focusing on tonal issues [5] [7]. There are some works on tone correctness improvements for Thai [4] [14]. However, the experiment focused on improvement methods; the subjects were requested to decide whether the syllables had the same tones as the given texts or not. We believe that there is no special design to distinguish tones in the same context.

This paper proposes perceptual evaluations focusing on tonal aspects of a Vietnamese TTS system. The first evaluation is a MOS test to assure the general quality of the system. The second is a syllable, tone and phoneme intelligibility test. The final test is devoted to tone intelligibility including some special designs for tone perception in the same context. These evaluations were conducted on the synthetic speech of VTed using a natural speech reference.

The rest of this paper is organized as follows. Section 2 presents the development of an HMM-based TTS system for Vietnamese and the implementation of VTed for Hanoi Vietnamese. Section 3 presents and discusses experimental results for three kinds of perceptual evaluations of VTed. The final section gives conclusions and presents future works.

2 Development

2.1 System architecture
The proposed architecture of an HMM-based TTS system for Vietnamese language is illustrated in Figure 1. There are three parts to this architecture: Natural language processing (NLP), Training, and Synthesis.

![Figure 1: Architecture of the HMM-based TTS system for Vietnamese.](image-url)
module includes both tonal and syntactic prosody modeling, based on tonal and syntactic information from previous modules. The Feature Processing module inputs all information from previous modules; process them to build a set of features including contextual factors in phoneme, syllable, word, phrase, and utterance level. Factors exist related to tones such as tone type, tone features at different levels.

The Training part uses two main inputs to produce a trained voice using HMM and EM algorithms: (i) Speech parameters including spectral (mel-cepstrum) and excitation parameters, which are extracted from the audio corpus and (ii) Context-based features (extracted from the text corpus) aligned with labels (automatic labeling of the audio corpus).

In the Synthesis part, context-based features are used to produce a sequence of speech parameters in such a way that its output probability for the HMM is maximized. High-quality synthesized speech is obtained using these speech parameters and a vocoder.

2.2 Vietnamese phonetic and phonology

2.2.1 Vietnamese syllable structure

The structure of Vietnamese syllables has been the subject of strong debate. The study in [15] concluded that there are four parts to the structure: initial, nucleus, ending, and tone (no rhyme). He also argued that the medial should be considered as a ‘semi-vowel’ instead of a main part in the structure. However, [16] and [17] preferred the hierarchical structure and affirmed the important role of rhyme in the structure of Vietnamese syllables. The authors in [16] did not take tones into account in the structure and they argued that medial (/w/) is in nucleus. For the tone, [17] assumed that the tone is a part of the whole syllable. The study in [18] performed a perception test using a Diagnosis Rhyme Test (DRT) method and concluded that the initial does not take part in the construction of the tone, which means that the Vietnamese tone affects only the rhyme of the syllable.

With all the above analysis, we adopted the hierarchical structure with two main parts of a syllable: An initial consonant and a rhyme. A tone is one part of rhyme with three other elements: medial, nucleus, and ending. The nucleus and tone are compulsory while others are optional.

2.2.2 Vietnamese phonological system

Table 1: Vietnamese initial consonants

<table>
<thead>
<tr>
<th>Place of articulation</th>
<th>Manner of articulation</th>
<th>Bi-</th>
<th>Labio-</th>
<th>Alveo-</th>
<th>Palato-</th>
<th>Ve-</th>
<th>Glo-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td></td>
<td>m</td>
<td>n</td>
<td>p</td>
<td>t</td>
<td>c</td>
<td>k</td>
</tr>
<tr>
<td>Plosive</td>
<td>Aspirated</td>
<td>p’</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Un-aspirated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voiced</td>
<td>Voiceless</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voiced</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant (Liquid)</td>
<td>Voiceless</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Approximant (Liquid)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 presents our conclusion for Hanoi Vietnamese initial consonants. In this dialect, orthographic ch /c/ and tr /t/ do not have their own place of articulation and are pronounced as /k/, /z/ and gi, r /j/, x /s/ and s /s/ are produced as /j/ /w/ (semi-vowels). Following back rounded vowels /u o ɔ/ the velar stops /k N/ are produced as doubly articulated labial-velars /kp ɲm/ [21]. Hanoi Vietnamese distinguishes nine short vowels /a ɔ u ɔ̌ ə ɛ ɪ ʊ o/ four short diphthongs /iə uə ɤ ɔ̌/ and three falling diphthongs /ie ue uo/ [17], illustrated in Table 2.

Table 2: Vietnamese vowels/diphthongs

<table>
<thead>
<tr>
<th></th>
<th>Position</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unrounded</td>
</tr>
<tr>
<td>Close (High vowel)</td>
<td>i ie</td>
<td>u u ɪ</td>
<td>u u o ʊ</td>
<td></td>
</tr>
<tr>
<td>Close-mid</td>
<td>e e</td>
<td>y y</td>
<td>o o</td>
<td></td>
</tr>
<tr>
<td>Open-mid</td>
<td>ε ε</td>
<td>ι ι</td>
<td>ι ι</td>
<td></td>
</tr>
<tr>
<td>Open (Low vowel)</td>
<td>a á</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2.3 Vietnamese tones

Northern Vietnamese is known as a tonal language having six different lexical tones. These are: level (1), falling (2), broken (3), curve (4), rising (5), and drop (6) tones. The study of [22] [23] confirms that voice quality is a robust correlate of tone in Hanoi Vietnamese, showing less variability than F0 across reading conditions. The experiment in [23] warrants the conclusion that rising (5b) and drop (6b) tones of syllables ending in /p, /t, /k/ (checked syllables) are not glottalized, either in final or non-final position. The work on oral flow [24] brings out a clear difference between these two sets of rhymes: tone 6a (drop tone in unchecked syllables) has low oral airflow; tone 5b and 6b have relatively high oral airflow, getting close to the range of breathy voice.

Table 3: Characteristics of 8 Vietnamese tones

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Register</th>
<th>Duration</th>
<th>F0 contour</th>
<th>Phonation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Level</td>
<td>High (+)</td>
<td>Long (+)</td>
<td>Level</td>
<td>Modal voice</td>
</tr>
<tr>
<td>2</td>
<td>Falling</td>
<td>Low (-)</td>
<td>Long (+)</td>
<td>Falling</td>
<td>Breathy voice</td>
</tr>
<tr>
<td>3</td>
<td>Broken</td>
<td>High (+)</td>
<td>Long (+)</td>
<td>Falling-Rising</td>
<td>Glottalization</td>
</tr>
<tr>
<td>4</td>
<td>Curve</td>
<td>Low (-)</td>
<td>Long (+)</td>
<td>Rising</td>
<td>Harsh voice</td>
</tr>
<tr>
<td>5a</td>
<td>Rising</td>
<td>High (+)</td>
<td>Long (+)</td>
<td>Rising</td>
<td>Modal voice</td>
</tr>
<tr>
<td>5b</td>
<td>Rising</td>
<td>High (+)</td>
<td>Short (+)</td>
<td>Rising</td>
<td>Tense voice</td>
</tr>
<tr>
<td>6a</td>
<td>Drop</td>
<td>Low (-)</td>
<td>Short (+)</td>
<td>Dropping</td>
<td>Glottalization</td>
</tr>
<tr>
<td>6b</td>
<td>Drop</td>
<td>Low (-)</td>
<td>Short (+)</td>
<td>Dropping</td>
<td>Tense voice</td>
</tr>
</tbody>
</table>

Therefore, it could be said that Vietnamese has a six-tone paradigm for sonorant-final syllables, and a two-tone paradigm for obstruent-final syllables [23], summarized in Table 3.

2.2.4 Vietnamese allophones

As presented in subsection 2.2.2, there are in total 21 consonants, 13 vowels, three diphthongs and two semi-vowels in Hanoi Vietnamese. However, if we also consider the function of each phoneme in the syllable, there are 19 allophones for initial, one allophone for medial, 18 allophones for nucleus and 10 allophones for ending (including two semi-vowels).

Table 4: Methods to build the allophone set for Vietnamese

<table>
<thead>
<tr>
<th>Function of phone</th>
<th>Embed tone?</th>
<th>Results</th>
<th>Allophones #</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>21 consonants, 13 vowels, 3 diphthongs &amp; 2 semi-vowels</td>
<td>39</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>19 initials, 1 medials, 18 nucleus and 10 endings</td>
<td>48</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes, in nucleus</td>
<td>19 initials, 1 medials, 18 nucleus and 10 endings</td>
<td>138</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes, in all elements of the rhyme</td>
<td>19 initials, 1x6 medials, 18x6 nucleus and 10x6 endings</td>
<td>193</td>
</tr>
</tbody>
</table>

In fact, based on the discussion of syllable structure, the tone is produced in parallel with each phoneme in the rhyme during articulation. It could be said that the same phoneme in the same rhyme bearing different tones is unalike. HMM-based speech synthesis models parameter sequences, i.e.
sequences of allophones. Therefore, beside tonal factors, it is wiser to embed tones inside suitable phonemes in each syllable, rather than considering tones as phonemes. Table 4 summarizes different ways to take tones into account to build up the allophone set. After carrying out a preliminary evaluation for each method, we adopted the last method because it produced the best quality for tone synthesis.

2.3 Implementation of VTed

There are several platforms that can be used to develop an HMM-based TTS system [9]. We have chosen the Mary TTS platform to build an HMM-based TTS system for Hanoi Vietnamese (VTed) because of its ease of use, expandability, and a high quality vocoder.

2.3.1 Implementation of VTed

We have built an NLP part integrated with the other modules in Mary TTS to establish an automatic training and synthesis process. All modules are implemented following the architecture in Figure 1. We adopted the methods in [25] to build the Word Segmentation and in [26] to build the POS Tagging module. For the Text Normalization module, we integrated our previous work [27] into the NLP part. The Proxody Modeling module is derived from the ToBI model to obtain features related to the ToBI end tones of syllables and phrases. The G2P & Tone Extraction module is designed specially to easily change different allophone units. The Feature Processing module and Feature Label Alignment module are extensions of the existing modules in Mary TTS that are suitable for Vietnamese.

Other modules were reused from the Mary TTS platform as the following main points [28]: (a) The Labeling module: Using the eHMM tool of Festvox to label audio corpus automatically, (b) The Parameter Extraction module: Using the SPTK and Snack tools for extracting features as in the original HTS scripts, (c) The HMM-based Training module: Using a version of the speaker dependent training scripts provided by HTS that (i) use context features predicted by the NLP part, (ii) include global variance [29], (iii) compose training data from mel-generalised cepstrum (mgc) [30], log F0 and strength files, and (iv) include features (band pass voicing strengths) for generation of mixed excitation [31]. (d) The Parameter Generation module and HMM-based Synthesis module: a new HMM-based synthesizer ported to Java from the HTS with the extension of a mixed excitation vocoder.

2.3.2 VN-Voice Training

There are 84.31% of words composing of at least two syllables [25]. Based on the syllable structure in subsection 2.2.1, there are four elements at the phoneme-level. Therefore, we used 5-state left-to-right HMMs to capture all the elements of syllables and boundaries of syllables and words.

The training is automatically carried out with a corpus of ~92% from 630 sentences from our existing corpus, VNSpeechCorpus, while the remainder (8%) are used in the evaluation phase. These sentences are recorded by a Vietnamese female broadcaster from Hanoi at 48 kHz and 16 bits per sample. Total duration of all sentences is about 37 minutes.

3 Perceptual evaluation

The perceptual evaluations included the assessment of general MOS, intelligibility of utterances and its elements, and tone perception in context. These evaluations are carried out with VTed and a natural speech reference, presented in random order. 18 subjects (9 females, 9 males) participated in the tests. All subjects are from the North of Vietnam, living for a long time in Hanoi. Participants were 20-35 years old and reported normal hearing and vision.

3.1 Evaluation of general quality

Subjects were asked to score “5-Excellent, 4-Good, 3-Fair, 2-Poor and 1-Bad” for overall impression after listening to an utterance. There were 48 sentences in the test corpus (8% of VNSpeechCorpus). For the sake of comparison, this test was also carried out on our previous TTS system adopting non-uniformed unit-selection synthesis - HoaSung [13] using the same training corpus (92% of VNSpeechCorpus).

A two-factorial ANOVA was run on the results. The two factors were the TTS system (3 levels) and the Sentence (48 levels). All factors and their interactions have highly significant effect (p<0.001); meanwhile the TTS system factor alone explains an important part of the variance (partial η2=0.63), while the Sentence factor and the interaction explain only about 15% each. A post-hoc Tukey test shows that each TTS system received significantly different mean scores. The experiment results plotted in Figure 2 show that the sound quality of VTed is rather good (0.81 point higher than HoaSung), but still clearly distinguishable 1.21 point lower) from natural speech.

3.2 Evaluation of utterance-level intelligibility

Subjects were asked to write down texts of utterances they heard. They could listen to samples one, two, or three times. 36 sentences with a length of 8 to 20 syllables were extracted from newspapers.

The approximate string matching [32] and the Damerau-Levenshtein algorithm is adopted to compute the distance between two strings with some improvements in dynamic programming. The error rates for syllable, tone, and phoneme is calculated based on the metric proposed in [33].
The error rates of VTed and natural speech are illustrated in Figure 3. VTed diverges from natural speech from 0.2% - 0.9%.

3.3 Evaluation of tone intelligibility

3.3.1 Stimuli and paradigm

In the tone intelligibility test, groups of meaningful sentences with the same syllables, diverging only for one tone, were prepared. As syllables with a systematic variation of tones on the same syllable are not always found, it is necessary to use proper names in some cases. Subjects were asked to choose the most likely syllable they heard among a group of syllables bearing different tones in an utterance, for example:

- Ở đây có buôn bán … không? (Do you sell … here?)
  - dể (goats – level tone 1)
  - dể (easily – broken tone 3)
  - dể (crickets – rising tone 5a)

- Mỗi tối, bác sĩ … thường đến hỏi thăm các bệnh nhân (Every evening, doctor … usually visits her patients)
  - Thụy (a person name with curve tone 4)
  - Thụy (a person name with falling tone 2)
  - Thụy (a person name with rising tone 5a)
  - Thụy (a person name with drop tone 6a)

For each utterance, we also included “Not like all above options” in the answer list. The test contains 129 sentences in 40 groups with a good balance of 9-12 examples for each tone pair. In this test, subjects could decide if they needed to repeat the sample one more times after the first listening.

3.3.2 Experiment results

The tone intelligibility rates are illustrated in Figure 4. The results are from 15-18% higher for utterances needing a single listening (once) than for those needing two listening repetitions (twice). The synthetic speech shows about 22.5% correct identification rate, lower than the natural speech in the global result. The proportion of “Not like all above options” selection is minor, from 0.0 to 0.7% for natural speech, and from 1.0 to 3.4% for VTed.

The corrected rates by tone types in Figure 5 show that all tones are nearly 100% perceived in the same context, except the falling tone 2. It appears that this tone is the most difficult tone for both natural and synthetic speech.

The result of VTed can be divided into three groups, regarding the perception test results. In the first group, 100% of tones are correctly perceived, as for natural speech. This group contains the rising tone. The second one, with the perception rates from 18 to 22% lower than natural speech, contains the level tone 1, drop tone 6b (around 87% correct perception) and the rising and drop tones 5a, 5b (around 80% correct perception). The last group, with the perception rates from 29% to 37% lower than natural speech, contains the broken tone 3, curve tone 4 and falling tone 2 (around 62% correct perception).

![Figure 4: Correct rates of tone intelligibility.](image)

Figure 5: Correct rates by tone types of tone intelligibility.

4 Conclusions and perspectives

Based on our study on Vietnamese phonetic and phonological features and design of a TTS system for Vietnamese, an HMM-based TTS system for Hanoi Vietnamese (VTed) is developed. The allophone set is built at the phoneme-level regarding the tone type and the function of phoneme in rhyme. Context-based features, including tone features, are extracted automatically for both training and synthesis phase in VTed. Three kinds of tests are conducted to assess the quality of VTed. In the MOS Test, the score of VTed was 1.21 points lower than natural speech, while the error rate of utterance-level intelligibility was only from 0.2% - 0.9% higher than the natural speech. These results show that VTed can produce rather good speech. In the tone intelligibility test, subjects were asked to recognize the same syllable bearing different tones in the same carrier sentence. The correct identification rate of VTed in this test was 23% lower than for natural speech on average (with 100% accuracy for the rising tone 5b and a 18% - 37% difference between synthetic and natural speech depending on the tone types).

These results show the need for more work to improve the quality of an HMM-based Vietnamese TTS system. Corpus design should consider the allophone unit (i.e. initial consonants and rhyme instead of phoneme-level) and the balance of allophones/tones, allophones in (tonal) context. More efforts are needed on tonal and/or syntactic prosody modeling in the NLP part, instead of the ToBI model. Due to the phenomena of glottalization in Vietnamese, this task may study the influence of not only F0 evolution, but also intensity and duration on tones in context. In addition, tone perception can be improved by using tone features to build design a decision-tree structure (as the work for Thai [4]).

5 Acknowledgements

We would like to thank Lê Quang Thắng and Nguyễn Thị Lan for their supports in recording and conducting tests at MICA.
6 References


