INTERSPEECH 2009 BRIGHTON

Cross-language F0 Modeling for Under-resourced Tonal Languages: A Case Study on Thai-Mandarin

Vataya Boonpiam, Anocha Rugchatjaroen, Chai Wutiwiwatchai

Human Language Technology Laboratory,
National Electronics and Computer Technology Center (NECTEC)
Patumthani, Thailand
{vataya.boo,anocha.rug,chai.wut}@nectec.or.th

Abstract

This paper proposed a novel method for F0 modeling in under-resourced tonal languages. Conventional statistical models require large training data which are deficient in many languages. In tonal languages, different syllabic tones are represented by different F0 shapes, some of them are similar across languages. With cross-language F0 contour mapping, we can augment the F0 model of one under-resourced language with corpora from another rich-resourced language. A case study on Thai HMM-based F0 modeling with a Mandarin corpus is explored. Comparing to baseline systems without cross-language resources, over 7% relative reduction of RMSE and significant improvement of MOS are obtained.
Index Terms: F0 modeling, cross language, tonal language

10.21437/Interspeech.2009-184

1. Introduction

F0 modeling is important for improving the naturalness of speech produced by a text-to-speech synthesizer. Algorithms proposed for this task vary depending on many factors such as unit size (e.g. syllables or phrases), characteristics of languages (e.g. intonational or tonal languages), and the type of F0 models (e.g. parametric and non-parametric). Conventional approaches often require a speech corpus whether for analyzing or for training the model. Parametric models such as Fujisaki [1] and Tilt models [2] represent an F0 contour with a set of equations. Parameters of these equations are analyzed from speech corpora using a machine-learning approach such as CART. An F0 contour can also be estimated in a non-parametric manner by selecting it directly from a speech corpus [3]. These F0 modeling approaches suffer from insufficient speech resources when applying to under-resourced languages. Without a speech corpus, F0 contours might be simply produced by rules. However, the result is known to be unnatural.

In tonal languages, an important cue for predicting F0 contours is a syllabic tone. Each tone has a distinct F0 shape over a duration axis. From our observation, we found that some F0 shapes are somewhat similar across tonal languages. For examples, a rising tone where F0 keeps increasing along the duration appears in several tonal languages such as Thai, Mandarin, and Vietnamese. This motivated us to explore the possibility of cross-language F0 modeling where similar tones across languages are mapped and an F0 model of an under-resourced language is augmented with speech corpora from a rich-resourced language. Cross-language approaches have been proposed to resolve the problem of resource sparseness but mainly for spectrum modeling [4]. To the best of our knowledge, this work is among the first that applies a cross-language approach to the problem of F0 modeling.

In this paper, we applied the cross-language approach to improve the quality of F0 modeling for Thai whose resources are limited. With tones mapped to Mandarin tones, an available large Mandarin corpus is expected to help enhancing Thai F0 modeling. HMM is shown to be suited to our idea since it provides an efficient way to adapt to specific data. In our approach, we first train an HMM for each Thai tone using the Mandarin speech corpus. This initial model is then adapted using a small set of Thai speech utterances. The proposed model was evaluated both objectively and subjectively with a comparison to conventional F0 models using none of cross-language resources.

This paper is organized as follows. The next section describes the relationship between F0 contours and tones in tonal languages. The characteristic of Thai tones, our case study, is also given. Section 3 reviews existing F0 modeling techniques for Thai and explains our proposed cross-language approach. Experiments and results are given in Section 4. Section 5 gives a discussion and concludes this paper.

2. F0 Contours in Tonal Languages

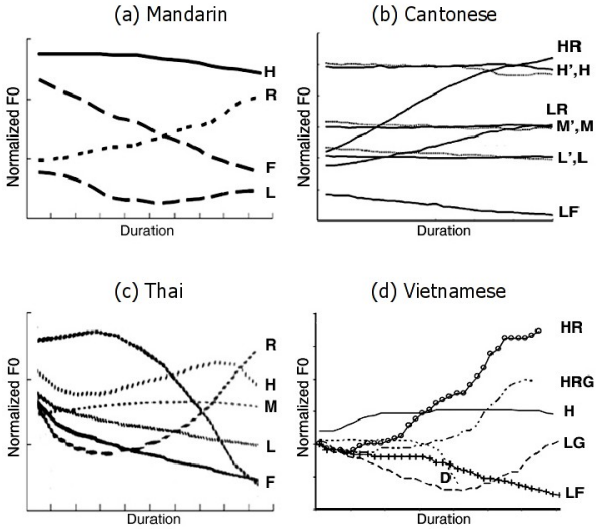

Figure 1: Examples of syllable-unit F0 contours in four tonal languages; (a) Mandarin [6], (b) Cantonese [7], (c) Thai [8], and (d) Vietnamese [9].

Tone is a very crucial suprasegmental feature in a tonal language as it affects word meaning. Tones are characterized by the shapes of their F0 contours over syllable duration. Figure 1 illustrates some examples of syllable-unit F0 contours in four tonal languages; Mandarin, Cantonese, Thai, and Vietnamese. Each F0 curve in each graph represents one tone type in that particular language.

Tones are often named according to their F0 shapes and levels such as rising (R), falling (F), low (L), medium (M), and high (H). Some tones are a sequential combination of these events such as LR and HR. Fujisaki et al. [5] introduced a phonological structure for expressing tone systems by using a pair of tonal commands for each tone where each command can be either positive or negative command. Figure 2 represents phonological structures of the four tonal languages, where each circle indicates a tone. We can see from these structures that although some tones are similarly named, their actual curves can be varied. For instance, H tones in Thai and Mandarin are different by their F0 levels.

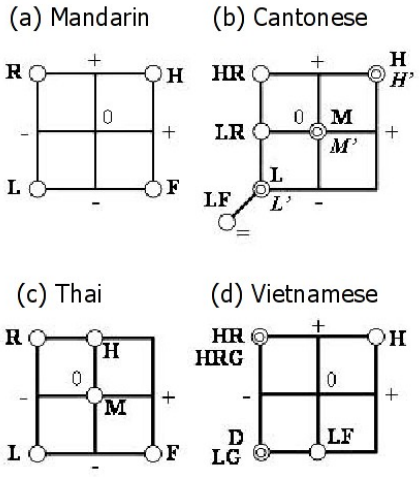

Figure 2: Phonological structures of four tonal languages [5]; (a) Mandarin, (b) Cantonese, (c) Thai, and (d) Vietnamese.

3. F0 Modeling for Thai

In this paper, we focus on F0 modeling for Thai. Existing F0 models are described in Section 3.1 while our proposed cross-language model is described in Section 3.2.

3.1. Corpus-based F0 modeling

Given a speech corpus in a particular language, there are a number of approaches that can be used to model F0 contours. One simple model is to select appropriate F0 units from the given corpus and apply them to synthesized speech [3]. The quality of the synthesized speech then strongly depends on a unit selection algorithm as well as the sufficiency and coverage of speech data. Several parametric models can also be applied such as the Fujisaki [1] and the Tilt [2] models. These models describe the F0 contour of a particular unit by a set of equations. In synthesizing stage, the parameters of these equations are predicted from contextual information of the input text using a machine-learning algorithm such as CART. Examples of contextual information are phonetic, syllabic, word or phrasal information and prosodic information such as syllable duration. In an HMM-based speech synthesis framework [10], F0 can also be modeled simultaneously with the spectral parameters in frame-based HMM units.

Regarding Thai, a recently investigated model is called T-Tilt [11], which is modified from the Tilt model [2] to better fit to tonal languages. Similar to [2], T-Tilt parameters can be estimated using CART models trained from a number of contextual features such as tonal context and part-of-speech.

3.2. Cross-language F0 modeling using HMM

It is obvious that speech resources are very crucial for all above mentioned methods. For languages with resource scarcity, it is very interesting to explore a possibility to borrow resources from other languages. The relationship between tones and F0 contours described in Section 2 provides a feasible way to map similar tones among different tonal languages. This mapping enables us to model F0 in one language with corpora from another language. The following subsections explain in detail how this idea is applied to our case study on Thai F0 modeling using Mandarin speech data.

3.2.1. Tone mapping between Thai and Mandarin

To model F0 in Thai using our cross-language approach, we need to find another language whose tone characteristics are close to Thai. The phonological-structure plot as illustrated in the Figure 2 is one good source for observing language similarity in terms of tone systems. In our case study, Mandarin is a suitable choice since its tone system is similar to Thai's plus we can obtain a large amount of Mandarin data. Based on the idea of tonal characteristic similarity, it is possible to find a suitable language for other targeted languages.

Mandarin consists of 4 tones as shown in the Figure 1(a); rising (R), falling (F), low (L), and high (H), plus an additional tone called "neutral", which is observed on non-tonal syllables. From these F0 contours and the ones for Thai tones shown in the Figure 1(c), we can create a mapping between Mandarin tones and Thai tones as shown in Table 1. It should be noted here that the neutral tone in Mandarin is actually observed to be non-tonal unit often appeared in short syllables in natural speech, whereas the medium tone in Thai appears to be a flat shape without any positive or negative tonal command. Although these two tones are different in some senses, they do have some overlapped characteristics and might be possibly mapped to each other.

Table 1: Mandarin and Thai tone mapping.

Thai	Mandarin
Medium (M)	Neutral
Low (L)	Low (L)
Falling (F)	Falling (F)
High (H)	High (H)
Rising (R)	Rising (R)

3.2.2. F0 modeling using HMM

HMM is a suitable F0 modeling engine in our case due to its ability to adapt with targeted speech data. HMMs are initialized for each Thai tone using the Mandarin speech database whose tones are mapped to Thai. Features used to train the HMMs are simply a frame-based F0 value augmented by its first and second derivatives. To model in a context dependent fashion, a clustering tree is built by considering tone-related questions as shown by the HTK toolkit format in Figure 3. Questions are defined to cluster tonal HMM states into groups according to their characteristics of tonal context. In Thai, two broad groups of tone, dynamic or static, can be applied. The dynamic tones include rising and falling whereas the static tones include high, medium, and low.

```
QS "Left_Dynamic"   { Falling-*,Rising-* }
QS "Left_Static"    { High-*,Low-*,Med-* }
QS "Left_Falling"   { Falling-* }
QS "Left_High"  { High-* }
.....
QS "Left_Silence" { Sil-* }
QS "Right_Dynamic" { *+Falling,*+Rising }
.....
```

Figure 3: Tone-related questions in the HMM state clustering tree.

Coarticulation phenomena of connected tones might be considered similar across languages. For instance, the ending of a rising tone might be reduced from its normal value if it is followed by a low tone. This motivates us to make use of rich tonal context in the large Mandarin corpus for Thai F0 modeling. However, local F0 contours within the syllable unit could be language dependent. For examples, the F0 contour of rising tones in Mandarin and Thai can be partially different. With a high perceptual sensitivity of native tonal-language people over distorted tones, it is necessary to modify local F0 shapes of Mandarin tones by using F0 shapes of Thai tones. This can be done easily in the HMM space where F0 values are encoded with Gaussian mixture parameters. Given a small adaptation set, several algorithms such as MLLR or MAP can be applied to modify the parameters. Re-estimating the initialized parameters given the data set is also another way to use the adaptation data. These adaptation methods require only a small adaptation set of Thai utterances.

In the synthesizing stage, context-dependent HMMs are chosen according to an input tone sequence. It is noted that an HMM can always be constructed for any unseen tonal context by utilizing the clustering tree. Once we get the HMMs of the given tone sequence, an F0 contour can be generated using algorithms such as that proposed by Tokuda et al. [10].

4. Experiments

In our study, Thai is a targeted language that we would like to model F0 while Mandarin is a matched language whose tone system are close to Thai. This section explains the experimental set up and results from both objective and subjective tests.

4.1. Experimental data

Table 2: Details of three sets of experimental data.

Characteristics	ChTR	ThAD	ThTS
Duration (min.)	216.0	15.4	36.1
No. of utterances	4,000	100	350
No. of syllables	51,269	3,017	5,466
Speakers	1 female	1 female	5 males, 5 females
Recording cond.	Clean	Clean	Clean
Purpose in this work	Initializing tonal HMMs	Adapting tonal HMMs	Evaluating models

Three sets of speech utterances, denoted as “ChTR”, “ThAD”, and “ThTS”, are used in our study. The details of these data sets are given in Table 2. It is noted that the experimental data are designed according to a practical situation where only a small amount of speech data is available in a targeted language (Thai) while a much larger amount of speech data is available in a matched language (Mandarin). , To eliminate the effect of speaker dependency when evaluating the proposed cross-language F0 model, various speakers are included in the evaluation set. However, these speakers are not included in the training and the adaptation set.

4.2. Objective test

The proposed cross-language F0 model, denoted as “HMM-Cross” consists of 5 tonal context-dependent HMMs one for each Thai tone. Each HMM is composed of 5 states with 1 Gaussian mixture per state. HMM states are tied using the clustering tree shown in the Figure 3. F0 values and their first and second derivatives extracted from the ChTR utterances are used to train the initial tonal HMMs using the HTS toolkit [10]. These initial HMMs are then adapted with the ThAD set. Our preliminary experiment showed that the simple re-estimation approach was superior to the use of MLLR or MAP adaptation algorithm. Final F0 contours can be synthesized also by the HTS tool.

To observe the effectiveness of our proposed cross-language model, a comparison to several F0 models is conducted. With a speech corpus, one can adopt a unit-selection-based approach [3] for F0 contours selection. Longest matching algorithm is applied to the sequence of input tones when performing unit selection. Two unit-selection models are investigated. “UnitSel-Ch” is a model that applies unit selection on only the ChTR corpus while “UnitSel-Th” is a model that applies unit selection on only the small set of Thai utterances, the ThAD set.

Without any cross-language attempt, F0 can also be modeled by the HMM, The model is called “HMM-Ch” when trained by the ChTR set and called “HMM-Th” when trained by the ThAD set. The last model is the T-Tilt model described in Section 3.1. The T-Tilt model trained from the whole set of TSynC-1 [12] serves in this paper as a recent advanced F0 model for Thai.

Table 3: Comparative RMSE results of various F0 models.

Model	RMSE (Hz)
UnitSel-Ch	74.2
UnitSel-Th	41.5
HMM-Ch	41.1
HMM-Th	41.2
HMM-Cross	38.3
T-Tilt	34.9

Table 3 reports root mean square errors (RMSE) between F0 produced by each model mentioned above and F0 extracted from natural speech. It can be seen obviously that the unit-selection approach is language sensitive. One reason for this is that different languages might be largely different in terms of frequently occurred tone sequences, tonal coarticulation, as well as local F0 shapes. Encoding F0 values in HMMs enables the construction of unseen tone sequences with highly smoothed F0 contours. This results in an enormous reduction of RMSE when using the HMMs. Training the HMMs or performing the unit-selection on a Thai only corpus should be the most appropriate approach if there are sufficient speech data in Thai. However, this is not the case in our experiment. With data limitation, our proposed cross-language HMM approach appears to be effective as it can reduce the RMSE by at least 7% relatively comparing to either the HMM-Th or the UnitSel-Th approach. With a z-score hypothesis test, our proposed model appears to be always superior to HMM-Ch

and HMM-Th models at over 99.7% confidence level and to the UnitSel-Th model at over 95% confidence level.

4.3. Subjective test

A human perceptual test is necessary in proving the effectiveness of the proposed model. A subset of 20 utterances is selected from the ThTS set by balancing the length of utterance. This subset contains 314 syllables in total. F0 contours of each utterance are replaced by the F0 curves produced by each of the three models; HMM-Th, HMM-Cross, and T-Tilt. Eleven Thai native speakers were asked to listen and grade each utterance by the 1-5 (worst to best) mean opinion scoring (MOS) according to their satisfaction on the naturalness of perceptual tones.

Another subjective test was performed on a syllable basis rather than an utterance basis in order to extensively analyze the intelligibility of our F0 model. In this second test, listeners were given syllable transcriptions of test utterances. After listening to each test utterance, they were asked to mark a syllable that has an incorrect tone. A tone error rate (%) is the number of incorrect syllables divided by the total number of syllables in the test set. Table 4 reports results from both mentioned subjective tests.

Table 4: Subjective test results.

Model	Avg. MOS	Avg. tone error rate (%)
HMM-Th	1.93	27.3
HMM-Cross	3.53	8.0
T-Tilt	1.75	31.8

According to the Table 4, the traditional HMM and T-Tilt models produce somewhat equal performance, while our cross-language HMM model significantly enhances both the naturalness and the intelligibility of synthesized speech. With a t-score hypothesis test, the proposed model always outperforms HMM-Th and T-Tilt models with over 99.9% confidence level. This promises the high possibility to Thai F0 modeling using the proposed cross-language approach. It is however worth noting that, from our observation, native speakers in tonal languages are very sensitive to tonal distortion. On the utterance-basis MOS measure, listeners tend to degrade the whole utterance even when the distortion appears on only few syllables. As discussed in [11], T-Tilt parameters could well represent F0 contours but were difficult to predict from given textual information. This results in the low MOS score for the T-Tilt model.

5. Discussion and Conclusion

This paper showed a high possibility to perform cross-language F0 modeling among tonal languages. The cross-language approach for modeling F0 in one language was done by training HMMs by a speech corpus from another language with tone mapping. The HMMs could then be improved by adapting with only a small set of utterances in the targeted language. The case study on Thai F0 modeling using a large speech corpus from Mandarin and a small set of Thai utterances showed the promising performance in terms of both objective and subjective evaluations.

To the best of our knowledge, this work is among the first exploration on prosody mapping in cross-language speech synthesis. It has then opened further interesting issues including investigations on other language pairs to ensure the usefulness of the proposed model, F0 modeling using multilingual databases, linguistics-driven versus data-driven tone mapping, modeling using other refined units such as sub-syllables, and the extension to other prosodic features required in speech synthesis. These topics are important for enabling speech synthesis of under-resourced languages under the cross-language scheme.

6. Acknowledgements

We would like to thank I^2R, Singapore, in providing a Mandarin speech corpus, and all anonymous reviewers in giving us useful technical suggestions.

7. References

[1] H. Fujisaki, and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, Journal of the Acoustical Society of Japan (E), 5(4), 233-241, 1984.

[2] P. A. Taylor, “Analysis and synthesis of intonation using the Tilt model”, Journal of the Acoustical Society of America, Vol. 107, no.3, pp. 1697-1714, 2000.

[3] A. Raux, and A.Black, “A unit selection approach to F0 modeling and its application to emphasis”, Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 700-705, 2003.

[4] C. Nieuwoudt, and E. Botha, “Cross-language use of acoustic information for automatic speech recognition”, Speech Communication, Vol. 38, 101-113, 2002.

[5] H. Fujisaki, and W. Gu, “Phonological representation of tone systems of some tone languages based on the command-response model for F0 contour generation”, Proc. International Symposium on Tonal Aspects of Languages (TAL), pp. 59-62, 2006.

[6] Z. Liu, J. Shao, P. Zhang, Q. Zhao, Y. Yan, and J. Feng, “Real context model for tone recognition in Mandarin conversational telephone speech”, Proc. 3rd International Conference on Natural Computation, pp. 696-699, 2007.

[7] T. Lee, G. P. Kochanski, C. Shih, and Y. J. Li, “Modeling tones in continuous Cantonese speech”, Proc. International Conference on Spoken Language Processing (ICSLP), pp. 2401-2404, 2002.

[8] Panroj, P., “Acoustic characteristics of tones in Bangkok Thai: variation by age groups”, Thesis (M.A.), Chulalongkorn University, 1991.

[9] T. T. Do, and T. Takara, “Vietnamese text-to-speech system with precise tone generation”, Acoustical Science and Technology 25, 5, 347-353, 2004.

[10] K. Tokuda, T. Kobayashi, and S. Imai, “Speech parameter generation from HMM using dynamic features”, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 660-663, 1995.

[11] A. Thangthai, N. Thatphithakkul, C. Wutiwiwatchai, A. Rugchatjaroen, and S. Saychum, “T-Tilt: a modified Tilt model for F0 analysis and synthesis in tonal languages”, Proc. International Conference on Speech Science and Technology (INTERSPEECH), pp. 2270-2273, 2008.

[12] C. Hansakunbuntheung, V. Tesprasit, and V. Sornlertlamvanich, “Thai tagged speech corpus for speech synthesis”, Proc. International Conference on Speech Databases and Assessments (O- COCOSDA), pp. 97-104, 2003.