Cross-Gender and Cross-Dialect Tone Recognition for Vietnamese

Antje Schweitzer, Ngoc Thang Vu

Institute of Natural Language Processing, University of Stuttgart, Germany

{antje.schweitzer|thang.vu}@ims.uni-stuttgart.de

Abstract

We investigate tone recognition in Vietnamese across gender and dialects. In addition to well-known parameters such as single fundamental frequency (F0) values and energy features, we explore the impact of harmonicity on recognition accuracy, as well as that of the PaIntE parameters, which quantify the shape of the F0 contour over complete syllables instead of providing more local single values. Using these new features for tone recognition in the GlobalPhone database, we observe significant improvements of approx. 1% in recognition accuracy when adding harmonicity, and of another approx. 4% when adding the PaIntE parameters. Furthermore, we analyze the influence of gender and dialect on recognition accuracy. The results show that it is easier to recognize tones for female than for male speakers, and easier for the Northern dialect than for the Southern dialect. Moreover, we achieve reasonable results testing models across gender, while the performance drops strongly when testing across dialects.

Index Terms: tone recognition, PaIntE, Vietnamese

1. Introduction

With around 7,000 languages in the world, one of the main challenging tasks nowadays is to gain a better understanding of all their properties to protect culture heritage, especially when many languages are in danger of becoming extinct. To date only a small fraction of languages has received much attention from the research community, viz. languages which are spoken by a large number of speakers in countries of great economic potential. Other languages, with interesting and challenging properties for speech and language technology, are mostly under-investigated. Many of these are tonal languages. Indeed, with the exception of Mandarin Chinese, most other tonal languages in the world belong to the under-resourced languages, and Vietnamese is one of them. It belongs to the Austro-Asiatic language family and is spoken by more than 90 million speakers in South East Asia. As in other tonal languages, the meaning of Vietnamese words changes depending on their tone. Vietnamese has six lexical tones, and a word like *'ma'*, for instance, has six different meanings depending on the tone:*'ma' (ghost), 'mà' (but), 'má' (cheek or mother), 'mả' (tomb or grave), 'mã' (horse or code), 'mạ' (rice seedling).* Thus tone recognition is important in order to process tonal languages like Vietnamese.

However there are only a few previous studies regarding tone recognition in Vietnamese [1, 2]. Their common approach is to represent each speech frame of the signal as a vector and to use a Hidden Markov Model in combination with a Gaussian Mixture Model for classification. Another approach to tone recognition uses more global context, representing the whole unit (syllable or the whole word) as a vector, and then employs classification algorithms such as neural networks, support vector machines and decision tree techniques for recognition. This approach has not been used for Vietnamese tone recognition so far, but for other tonal languages such as Thai [3] and Mandarin Chinese [4]. In this paper, we follow the second approach, representing each syllable as a vector and exploring the usage of decision tree and bagging techniques for classification. We also investigate the impact of features in addition to parameters such as single fundamental frequency (F0) values and energy features: we use harmonicity as well as the PaIntE parameters [5, 6], which to date have mostly been used for modeling intonation in English and German, which are both intonation languages. We explore whether these parameters can be applied for tonal languages, and how they affect recognition accuracy.

Although there is an increasing number of studies on Vietnamese speech processing, systematic studies exploring the impact of gender or dialect on tone recognition accuracy are still missing. [1] reported results on a very small database, investigating the cross-gender effect with only two speakers for each gender, and only for Northern Vietnamese. In constrast, we analyze the influence of gender and dialect on recognition accuracy on a fairly large speech database, the GlobalPhone speech data, which contains speech data for both male and female speakers and for two dialects, Northern and Southern Vietnamese.

We will describe the Vietnamese language in section 2 and a short overview of the PaIntE parameters in section 3. Section 4 describes the GlobalPhone Vietnamese database and our experiments. In section 5 we present an investigation of tone recognition across gender and dialect, and we present a conclusion in section 6.

2. The Vietnamese Language

Vietnamese is a monosyllabic tonal language. Each Vietnamese syllable consists of a syllable onset, a nucleus, and a coda and is associated with a tone. For example, the word *'toán' (math)* is a combination of a syllable onset *'t'*, a nucleus *'oa'*, a coda *n* and a *'high rising'* tone. There are, however, cases in which syllable onset or coda are empty, e.g. the word *'ma' (ghost)* contains an onset *'m'*, a nucleus *'a'*, an empty coda and a *'mid level'* tone. There are six lexical tones: T1 (mid level), T2 (low falling), T3 (high rising), T4 (mid dipping-rising), T5 (high breaking-rising) and T6 (low falling constricted) in Vietnamese [7], which affect word meaning, i.e. six different tones on the same syllable may result in six different words. In addition to its F0 shape, T5 is characterized by a high degree of laryngealization [13]. A more recent study [8] claimed that Vietnamese has 8 tones from a linguistic perspective, since T3 and T6 can be split in two different variants depending on the syllable coda (with voiced phones, or with stop consonant). However since the variants do not affect the meaning of the word, we focus only on recognizing six tones. Table 1 lists all tones with their descriptions and examples. Figure 1 illustrates the shape of the F0 contour of each of the tones over the syllable.

Table 1: *Vietnamese tones, their descriptions and examples*

Tone	Description	Examples
T1	mid level	*ma (ghost)*
T2	low falling	*mà (but)*
T3	high rising	*má (cheek, mother)*
T4	mid dipping-rising	*mả (tomb, grave)*
T5	high breaking-rising	*mã (horse, code)*
T6	low falling constricted	*mạ (rice seedling)*

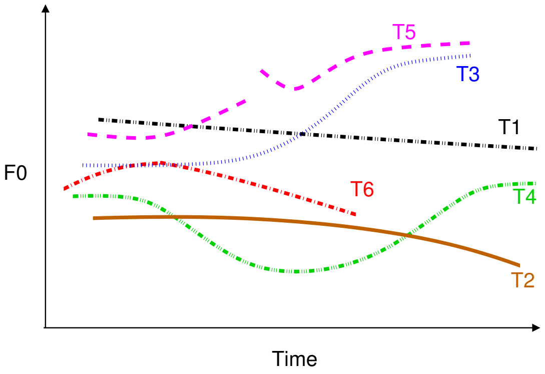

Figure 1: *F0 contour of six different lexical tones of Vietnamese language: T1 (mid level), T2 (low falling), T3 (high rising), T4 (mid dipping-rising), T5 (high breaking-rising) and T6 (low falling constricted), adapted from [7].*

3. The PaIntE Model

The PaIntE model [5, 6] was originally developed in the context of speech synthesis for intonation languages. It can be used to model the shape of the F0 contour in the vicinity of intonation events, hence the model's name, Parameterized Intonation Events, or PaIntE for short. The shape of the F0 contour is captured by six linguistically motivated parameters. In the "classical" PaIntE model, this contour is approximated over a window of three syllables. For the present purpose of modeling a tonal language, we are interested in the shape of F0 on single syllables only; thus we restricted the approximation window to one single syllable for the present experiment.

Mathematically PaIntE approximates the F0 contour by the following function, where x is syllable-normalized time.

$$f(x) = d - \frac{c_1}{1 + e^{-a_1(b-x)+\gamma}} - \frac{c_2}{1 + e^{-a_2(x-b)+\gamma}} \qquad (1)$$

This function yields a peak shape (Figure 2), where the *peak height* is determined by parameter d, its temporal location in the syllable by parameter b (*peak alignment*), the amplitudes of the rising and falling parts by parameters c_1 (*rise amplitude*) and c_2 (*fall amplitude*), and their steepness by parameters a_1 (*steepness of rise*) and a_2 (*steepness of fall*). Depending on the peak alignment and the amplitude parameters, this allows modeling falling or rising contours that involve some kind of peak. For cases without a peak, PaIntE employs two functions in which the second or the third term of the above equation are omitted, yielding a pure fall or a pure rise.

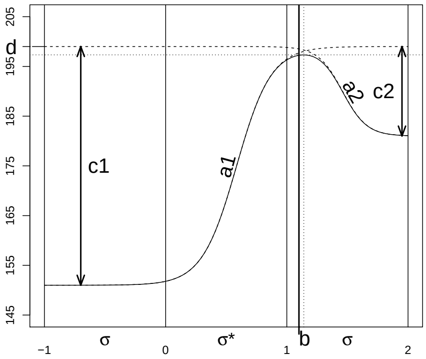

Figure 2: *Example PaIntE contour in a window of three syllables around the syllable of interest (σ*).*

Table 2: *Number of syllables with different tones*

Tone	#syllables
T1	42,850
T2	37,389
T3	48,340
T4	19,995
T5	9,830
T6	29,240

4. Data and Results

4.1. GlobalPhone Vietnamese Speech Data

We used the GlobalPhone (GP) Vietnamese speech database [9], which was collected as part of the multilingual GlobalPhone database [10]. Vietnamese native speakers were asked to read prompted sentences of newspaper articles. The resulting corpus consists of 25 hours of speech data by 140 speakers, from the cities of Hanoi and Ho Chi Minh City in Vietnam, as well as 20 native Vietnamese speakers living in Karlsruhe, Germany. Each speaker read between 50 and 200 utterances. In total, the corpus contains 22,112 utterances spoken by 90 male and 70 female speakers. All speech data were recorded with a headset microphone in a quiet environment. The Vietnamese GP data was originally developed for automatic speech recognition and was split in training, development and testing data. We used only the training data of the Vietnamese GP data in our experiments. Furthermore, we filtered all loan words consisting of more than one syllable, such as *pepsi, apple or google*. Table 2 lists the number of syllables we had for each tone.

4.2. Parameters

The PaIntE model is implemented as part of the Festival Speech Synthesis system [11]. We converted the Vietnamese speech data to Festival utterances, ran the PaIntE parametrization, and extracted the PaIntE parameters along with a number of other features available in Festival that we expected to be predictive of the tones. These included phonological features (first block in table 3), pitch values at the beginning, in the middle, and at the end of the vocalic part of each syllable (second block), as well as syllable duration.

These features were used in all following experiments. In addition, we used Praat [12] to calculate the energy within the vocalic parts, both over the complete vowel, and for each third of

Table 3: *Features used for tone recognition. The first block of features was used in every classifier and served as a baseline. Each of the following blocks corresponds to a class of features that we successively added to the feature set.*

phonological parameters	
onsettype	types of consonants in syllable onset
codatype	type of consonant in syllabe coda
vowelclass	height feature of vowel
onsetlength	number of consonants in onset
rhymelength	number of consonants in syllable rhyme
pitch features	
pitchbegin	pitch at beginning of vocalic part
pitchmid	pitch in middle of vocalic part
pitchend	pitch at end of vocalic part
deltapitch{1,2}	change from start to mid/mid to end
duration	
duration	duration of syllable

energy features	
allenergy	energy over complete syllable
energy{1,2,3}	energy in each third of the syllable
deltaenergy{1,2}	change in engergy between thirds

voicing feature	
voicing	number of voiced frames in vowel

spectral tilt	
tilt	spectral tilt

harmonicity	
harmonicity	harmonicity throughout syllable

PaIntE parameters	
p.a1, a1, n.a1	$a1$ for preceding, current, next syllable
p.a2, a2, n.a2	$a2$ for preceding, current, next syllable
p.b, b, n.b	b for preceding, current, next syllable
p.c1, c1, n.c1	$c1$ for preceding, current, next syllable
p.c2, c2, n.c2	$c2$ for preceding, current, next syllable
p.d, d, n.d	d for preceding, current, next syllable
approx. function	pure rise, pure fall, or peak shape

the vowel. We also used Praat to calculate the number of voiced frames in each vowel, and calculated long-term average spectra for which we retrieved the spectral slope to capture spectral tilt. Finally, we calculated harmonicity for each syllable to capture the laryngealization in tone T5 [13]. Table 3 gives an overview of all features we extracted. After feature extraction, we logged all pitch values as well as PaIntE parameter d (*peak height*), and standardized all pitch-related features by speaker. This is similar to the pitch normalization method suggested by [1], except that we normalize by speaker instead of by sentence. Energy values were logged and standardized by file.

4.3. Results

We used WEKA [14] to train classifiers for predicting tones given the features listed in table 3, using two learning schemes that had yielded best results in pitch accent classification for German [15, 16], viz. Random Forest, and Bagging, each using the default parameters suggested by WEKA. We started out using the phonological, pitch and duration features in the first block of table 3 as a baseline, and then succesively added the features classes in the next five blocks in that table, i.e. energy, voicing, spectral tilt, harmonicity, and finally the PaIntE parameters. We evaluated each classifier using 10-fold cross-validation as implemented in WEKA. Averaged accuracy rates across all 10 folds are indicated in table 4, with the input features indicated in the rows, and the two schemes in the columns. The third column indicates whether adding the class of features improved accuracy siginificantly at a level of $\alpha = 0.01$.

Table 4: *Averaged accuracy rates, obtained by 10-fold cross-validation, across the different feature sets, for Random Forest and Bagging classifiers. Baseline indicates the set consisting of phonological features, pitch, and duration. Each of the following lines gives the accuracy when adding another feature class (see the blocks in table 3). The last column indicates whether the improvement is significant at a level of* $\alpha = 0.01$.

Features	Rand. Forest	Bagging	
Baseline	57.68	65.54	*
+ energy	59.12	67.25	*
+ voicing	60.26	67.53	*
+ harmonicity	65.38	68.44	*
+ spectral tilt	65.60	68.55	n.s.
+ PaIntE	71.18	72.36	*

The results show that each of the classes of features we added improved recognition accuracy significantly, with the exception of spectral tilt. The strongest improvements come from adding harmonicity (improving accuracy by approx. 5% for the Random Forest classifier, but only by approx. 1% for the Bagging classifier), and from the PaIntE parameters (approx. 5.6% for Random Forest, and approx. 3.8% for Bagging). Thus the PaIntE parameters increase accuracy most consistently, and for both classifiers more than the harmonicity feature. The best accuracy is reached by the Bagging classifier when using the full featuere set including the PaIntE features.

5. Effects of gender and dialect

In this section, we investigate the influence of gender and dialect on Vietnamese tone recognition. The GlobalPhone database contains speech from female and male speakers from the Northern and the Southern dialect. We trained models separately for the two dialects and for both genders, i.e. for four different data sets. We again used Random Forest and Bagging classifiers in WEKA. For assessing gender and dialect-specific effects, we did not use 10-fold cross-validation in WEKA but split each data set into a test and a training part. This was (i) because we had different amounts of data in each data set and we wanted to ensure that each classifier is trained on a comparable amount of data and (ii) because we explicitly wanted to ensure that data from one and the same speaker are never in both the training and the test set. This is not guaranteed when using 10-fold cross-validation as provided by WEKA.

We had fewest data for Northern Vietnamese females. We split these data into two groups of speakers such that we obtained an approx. 80% to 20% split, while maximizing the number of speakers in the training data. The 80% training part corresponded to approx. 28,000 syllables from 20 speakers, the test part contained data from 3 speakers. For the three other groups, we selected comparable numbers of syllables (between

Table 5: *Accuracy rates for gender and dialect dependent models. The labels at the top of the columns indicate training/test data.*

	Northern Vietnamese	
Scheme	female/female	male/male
RandomForest	77.9299	71.0811
Bagging	77.7983	72.3287
	Southern Vietnamese	
Scheme	female/female	male/male
RandomForest	73.9863	64.2887
Bagging	74.6732	65.9603

28,000 and 29,000 syllables), yielding training sets from 18 to 22 speakers, leaving test sets with varying amounts of data from 8 to 11 speakers.

The accuracies obtained for these speaker and dialect dependent models are indicated in table 5. The best results can be obtained on female Northern Vietnamese, yielding approx. 78% accuracy for both schemes. However, both gender and dialect massively affect recognition accuracy: We observe differences of roughly 6% in accuracy (7% for Random Forest, 5.5% for Bagging) between female and male data for Northern Vietnamese, and of more than 9% in accuracy (10% for RandomForest, 8.5% for Bagging) between female and male data for Southern Vietnamese. Similarly, the results are better in general for Northern Vietnamese compared to Southern Vietnamese, with approx. 3.5% difference (4% for Random Forest, 3.5% for Bagging) between dialects for female data, and approx. 6.5% (7% for Random Forest, 6.5% for Bagging) for male data.

5.1. Gender

We investigate the effects of dialect and gender further by applying models across gender and across dialect. Table 6 indicates the accuracies when evaluating models trained on one gender and one dialect on test data of the other gender of the same dialect. The labels above each column indicate training/test data, e.g. female/male indicates accuracies for models trained on female data ("female models") when applied to male test data. Interestingly, female models perform worse on the corresponding male data than male models perform on female data (the accuracies in the left columns are in general lower than those in the right column), which seems to indicate at first glance that male models are better suited to be applied across gender.

However, they are by far not as good on female data as female models, thus on average (i.e. when testing on 50% male, 50% female data) female models would still be better than male models: for Northern Vietnamese for instance, we would expect an accuracy of around 78% for the female part (see cells female/female in top half of table 5), and of around 68-69% for the male part (see cells female/male in top half of table 6). Analogously for Southern Vietnamese: here we would get an accuracy of around 74% for the female part, and an accuracy of around 63-64% for the male part when using female models throughout, which on average yields roughly 69% accuracy. For both Northern and Southern Vietnamese, using male models on mixed data would yield lower accuracies.

5.2. Dialect

When applying models across dialects, the accuracy rates drop dramatically, as is evident from table 7. There is no clear pattern to observe—for female speakers, models trained on Northern data ("Northern models") perform better on Southern data than Southern models perform on Northern Data. However, for male speakers, we observe the opposite. The results are in general better for female Northern models and for male Southern models (around 55-57%), while male Northern models and female Southern models perform worst (around 50%). In summary, it is clearly not advisable to use classifiers across dialects.

Table 6: *Accuracy rates when applying models across gender. The labels at the top of the columns indicate training/test data.*

	Northern Vietnamese	
Scheme	female/male	male/female
RandomForest	68.26	69.99
Bagging	68.89	71.60
	Southern Vietnamese	
Scheme	female/male	male/female
RandomForest	62.63	69.65
Bagging	64.17	70.20

Table 7: *Accuracy rates when applying models across dialect. The labels at the top of the columns indicate training/test data.*

	Female	
Scheme	Northern/Southern	Southern/Northern
RandomForest	56.44	50.14
Bagging	55.86	51.63
	Male	
Scheme	Northern/Southern	Southern/Northern
RandomForest	50.44	57.03
Bagging	50.99	57.56

6. Conclusion

We explored new parameters for tone recognition in Vietnamese, showing that in addition to parameters used in earlier studies on tone recognition, such as normalized F0 end energy values and their deltas, new phonetically motivated parameters, viz. harmonicity and results from parameterizing the F0 contour using the PaIntE model, significantly improve tone recognition. Using female and male data from both Northern and Southern Vietnamese, we find that both harmonicity and the PaIntE parameters can lead to an overall improvement of approx. 5% each, leading to overall accuracies of up to 72%.

We further investigated the influence of gender and dialect on recognition accuracy. Here we reach even higher rates of up to 78% for female speakers of the Northern dialect. We have also shown that models generalize to some degree across gender. Although female models perform worse on male data than male models on female data, the female models' accuracy rates on same-gender data are so high that they would perform better on mixed data than male models (assuming poportions of 50% male vs. female data in a hypothetical mixed data set). In general, the gap in performance between female and male models is an interesting finding. Further investigation including a more detailed phonetic analysis could shed light on this aspect. Concerning cross-dialect models, the accuracies are much lower in general, suggesting that it is highly advisable to train separate models for the two dialects.

7. References

[1] H. Q. Nguyen, P. Nocera, E. Castelli, and V. L. Trinh, “Tone Recognition of Vietnamese Continuous Speech using Hidden Markov Model,” in *Second International Conference on Communications and Electronics*, 2008.

[2] P. N. Le, E. Ambikairajah, and E. H. Choi, “Improvement of Vietnamese tone classification using FM and MFCC features,” in *International Conference on Computing and Communication Technologies*, 2009.

[3] L. Tan, M. Karnjanadecha, and T. Khaorapapong, “A study of tone classification for continuous Thai speech recognition,” in *Proceedings of 8th International Conference on Spoken Language Processing.*, 2004.

[4] L. Wang, J. Zhang, B. Dong, and Y. Yan, “A SVM based tone recognition for Mandarin multi-syllable words,” in *Advances in Information Sciences and Service Sciences, 5(5)*, 2013.

[5] G. Möhler and A. Conkie, “Parametric modeling of intonation using vector quantization,” in *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 1998, pp. 311–316.

[6] G. Möhler, “Improvements of the PaIntE model for F0 parametrization,” Manuscript, 2001.

[7] V. L. Nguyen and J. A. Edmondson, “Tones and voice quality in modern northern Vietnamese: Instrumental case studies,” *Mon-Khmer Studies*, vol. 28, pp. 1–18, 1998.

[8] Q. Nguyen, N. Pham, and E. Castelli, “Shape vector characterization of Vietnamese tones and application to automatic recognition,” in *Proceedings of ASRU*, 2001.

[9] N. T. Vu and T. Schultz, “Vietnamese Large Vocabulary Continuous Speech Recognition,” in *Proceedings of ASRU*, 2009.

[10] T. Schultz, N. T. Vu, and T. Schlippe, “Globalphone: A multilingual text and speech database in 20 languages,” in *Proceedings of ICASSP*, 2013.

[11] Centre for Speech Technology Research, University of Edinburgh, “The Festival text-to-speech synthesis system,” [http://www.cstr.ed.ac.uk/projects/festival/].

[12] P. Boersma and D. Weenink, “Praat, a system for doing phonetics by computer [computer program],” 2016, version 6.0.14, retrieved 11 Feb. 2016. [Online]. Available: http://www.praat.org/

[13] D. D. Tran, E. Castelli, c. S. Jean-Fran V. L. Trinh, and X. H. Le, “Influence of F0 on Vietnamese syllable perception,” in *Proceedings of Interspeech (Lisbon, Portugal)*, 2005.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: An update,” *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[15] A. Schweitzer and B. Möbius, “Experiments on automatic prosodic labeling,” in *Proceedings of Interspeech (Brighton, UK)*, 2009, pp. 2515–2518.

[16] A. Schweitzer, *Production and Perception of Prosodic Events—Evidence from Corpus-based Experiments*. Doctoral dissertation, Universität Stuttgart, 2011.