AUTOMATIC LANGUAGE RECOGNITION WITH TONAL AND NON-TONAL LANGUAGE PRE-CLASSIFICATION

Liang Wang [1], *Eliathamby Ambikairajah* [2], *and Eric H.C. Choi* [3]

[1, 2]School of EE&Telecomm, the University of New South Wales
2032, Sydney, NSW, Australia
email: l.wang@student.unsw.edu.au, ambi@ee.unsw.edu.au
[3]ATP Research Laboratory, National ICT Australia
1435, Sydney, NSW, Australia
email: eric.choi@nicta.com.au

ABSTRACT

Parallel Phoneme Recognition followed by Language Modelling (PPRLM) systems currently provide state of the art language identification performance on conversational telephone speech. In this paper an innovative method for tonal and non-tonal language pre-classification by using prosodic information is reported. Our motivation is to improve recognition accuracy and save the amount of CPU run-time while handling large number of languages. Also, by incorporating different confidence measures into the traditional PPRLM framework, we propose an optimized language recognition system that can be applied in an open-set language recognition task. For a task of 12 target languages and 4 non-target languages, our results show that with the optimized pre-classification, Universal Background Phone Model confidence measuring and Witten-Bell discounting the system can achieve recognition accuracy rates of 77.9% for 30-sec speech segments and 49.2% for 10-sec speech segments.

1. INTRODUCTION

Human-machine interface applications increasingly leverage automatic language recognition techniques. These techniques are used in many applications, such as spoken language translation, call-routing, multi-lingual automatic speech recognition (ASR) and cross lingual ASR. Recent studies have explored a variety of methods that utilize different levels of speech features, including acoustic [1], phonotactic [2][3] and prosodic [4] features. So far the language recognition systems based on acoustic and phonotactic information produce the best results. The acoustic features are easier to obtain, but they are volatile as speaker or channel variations are present. The phonotactic features are believed to carry more discriminative information about the language, and to be more robust than acoustic features. The extraction of the phonotactic information, however, requires the speech to be labelled at a fine phone level for model training. This is a very time-consuming task. Language recognition systems based on prosodic information perform worse than those based on acoustic or phonotactic repertoire, the lack of an efficient way to model the prosodic characteristics being the primary reason. Systems based on prosodic information, however, are capable when dealing with a small number of target languages [5], or when the target languages need only be classified into broad categories [6].

Due to a lack of multilingual corpora that cover a variety of languages, recent language recognition systems [1][2][3][4] are only evaluated using up to 15 languages. Zissman [3] reported that by using the PPRLM system, the language identification error rate was 8% for a 45-s speech segment on a 3-language task. Using the same configuration, the error rate increased to 21% when evaluated using 11 languages. It should be noted that there are thousands of languages currently spoken in the world. Maintaining the recognition rate for much larger numbers of languages is a challenging problem.

Further, for current language recognition systems based on acoustic and phonotactic information, the computation time is largely dependent on the number of languages each language recognition system handles, increasing greatly as more languages are considered. Systems based on prosodic information, in contrast, generally require far less computation time.

We therefore propose to use prosodic information to perform a pre-classification before the final language recognition, where this final task will be based on acoustic or phonotactic information. After pre-classification, the total number of languages considered can be separated into two or three smaller subsets. The final language recognition rate will be increased by using a smaller language set, and computation time will be drastically reduced.

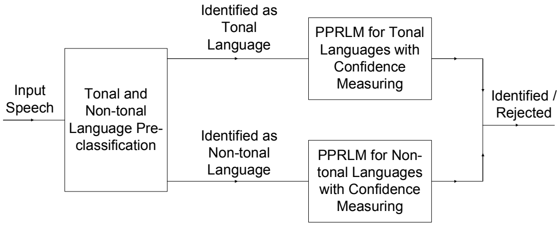
Input Speech
Tonal and Non-tonal Language Pre-classification
Identified as Tonal Language
Identified as Non-tonal Language
PPRLM for Tonal Languages with Confidence Measuring
PPRLM for Non-tonal Languages with Confidence Measuring
Identified / Rejected

Fig. 1. Overview of the Novel Language Recognition System with Tonal and Non-tonal Language Pre-classification

In this paper we propose a novel tonal and non-tonal language classification as a pre-classification for the language recognition task (Fig. 1). Based on our previous research, tonal and non-tonal language classification is performed by measuring the speed and level of pitch change. In order to keep the system robust, a PPRLM system is used for the final language recognition. We do this because our evaluation data comes from three different corpora, thus channel mismatching exists.

2. TONAL AND NON-TONAL LANGUAGE CLASSIFICATION

2.1. Characteristics of Pitch Information

In human languages, pitch is regarded as one of the important prosodic features that relate to phonation. It is obvious that the vocal folds can vibrate at different frequencies, and thus that vocal sounds can be produced at varying pitches [7][8]. Pitch and pitch changes are utilized in language in two distinct ways. On one hand, variations of pitch may be related to relatively long stretches of speech, many syllables in length, and correspond to relatively large grammatical units such as the sentence. Pitch variation used in this way is called intonation, and is used in all languages to express emphasis, contrast, and emotion. On the other hand, pitch variation can be used in short stretches of syllable length, such as in small grammatical units like words and morphemes. Pitch variation used in this way is called tone. Tonal languages are those that use tone to distinguish lexical meaning [6].

Based on our research into the pitch characteristics for tonal and non-tonal languages, we propose a novel technique for tonal and non-tonal language classification based on two pitch-change parameters, the speed and the level of pitch change.

2.2. Implementation of Tonal and Non-tonal Language Pre-classification

The pre-classification system has a number of input features, extracted from the input speech. Fig. 2 shows the construction of this system, with pitch information contributing 4 features to the classifier vector, in addition to a phoneme counter. Each component of the novel tonal and non-tonal language classification system is described in detail below.

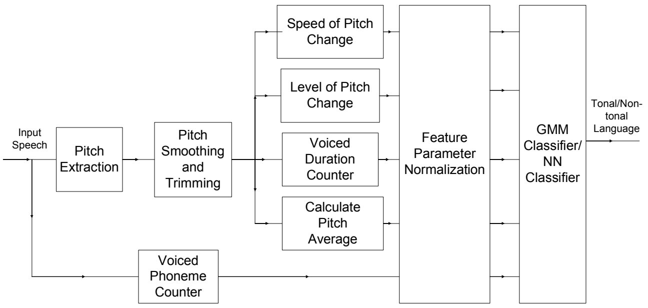

Fig. 2. The Block Diagram of the Novel Tonal and Non-tonal Language Classification System

i. Pitch extraction: The raw fundamental frequency (F0) contour is first automatically extracted based on an autocorrelation method from the input speech. In this experiment, the pitch values were extracted every 10 ms with a frame size of 40 ms.

ii. Pitch smoothing and trimming: Since pitch extraction from the speech signal is a difficult task to perform automatically and accurately, the resulting raw F0 contours often contain "spurious" pitch values during unvoiced speech segments, or "drop-out" in regions of voiced speech segments. Further, the lengthy stretches of aperiodicity due to creakiness would contribute to a phenomenon in pitch extraction called "double pulsing", in which the extracted pitch values are twice the actual value [9]. In order to get an accurate pitch contour, the pitch smoothing and trimming module is used after pitch extraction [6].

At first, all the voiced segments shorter than 100ms are removed. We assume these voice segments are produced by the spurious pitch values in regions of unvoiced speech segments, or they do not carry any actual meaning (these voiced segments may be caused by coughing, laughing, etc.). Following this, a trimming algorithm is used to smooth the F0 curves. The trimming algorithm compares the average pitch values (f_i) of a certain voiced segment against the average pitch value (F) of the whole utterance. The voiced segments will be kept for further processing only if:

$$C_1 * F < f_i < C_2 * F \qquad (1)$$

where C_1 and C_2 are two *a posteriori* thresholds. The trimming algorithm effectively eliminates sharp spikes in the pitch tracing often seen around nasal-vowel junctions. Finally, a 5th-order median filter is used to smooth the pitch again, and compensate for the "drop-out" during voiced speech segments.

iii. Speed of pitch change: Analysis of the speed of pitch change is performed on each voiced segment. Assume s_1, s_2, ..., s_j, ..., s_M stand for the voiced segments in a particular utterance, and f_1, f_2, ..., f_i, ..., f_N stand for the pitch values within a certain voiced segment in that speech utterance. The absolute value of the pitch variation within the voiced segment j is

$$pv_j = \sum_{i=1}^{N-1} | f_{i+1} - f_i | \qquad (2)$$

thus the total pitch change can be calculated by:

$$PV = \sum_{j=1}^{M} pv_j \qquad (3)$$

iv. Level of pitch change: Similar to the analysis of the pitch change speed, the pitch change level analysis is also performed for each voiced segment first. Let σp_j be the standard deviation of the pitch value of the j th voiced segment in the utterances. The pitch change level may then be measured by summing σp_j

$$\sigma_P = \sum_{j=1}^{M} \sigma p_j \tag{4}$$

Thus, the *speed* of pitch change is a measurement of the “local” pitch variation pattern, while the *level* of pitch change is used to measure the “global” pitch variation pattern.

v. Voiced duration counter: The voiced duration (VD) is estimated by counting the total number of voiced segments in the speech. For example, the total voiced sound duration of a speech utterance in our experiment was *VD*10ms,* as the pitch values were extracted every 10ms. The voiced duration is used to compensate for differences in speed between different speakers.

vi. Calculate pitch average: The average pitch value (AVE) is obtained by averaging the pitch value across the whole utterance. This average is used to normalize the speed and level of pitch change between different speakers. An example of this is to normalize between male and female speakers, as (generally) females have a higher average pitch than males.

vii. Voiced phoneme counter: The voiced phoneme count (VC) is defined as the number of voiced phonemes detected in a speech utterance. Voiced phonemes are identified by using a broad-phone-class recognizer. As the OGI-TS speech corpus [10] has already labelled six languages at the phonotactic level, we first translate the phonotactic-level labelling into phone-class-level labelling. We then build a HMM for each of the phone classes. Each phone-class is modelled by a 3-state HMM with 8 Gaussian mixtures.

The voiced phoneme count is used as a normalization factor, together with both the speed and level of pitch changes. Our motivation here is to examine the speed and level of pitch change for each phoneme symbol, as tonal variation can be present in each syllable in tonal, monosyllabic languages such as Mandarin.

viii. Feature parameters normalization: A normalization module is used to generate the feature parameters for the GMM classifier. Normalization of feature parameters is indispensable, reducing undesirable variation caused by speaker difference and other factors. In this study, PV and σ_P are normalized as:

$$P\hat{V} = \frac{PV}{AVE * VD} \tag{5}$$

$$\hat{\sigma}_P = \frac{\sigma_P}{AVE * VC} \tag{6}$$

So $P\hat{V}$ and $\hat{\sigma}_P$, together with *AVE*, *VD*, *VC* are the five feature parameters that are fed into the GMM classifier to perform the final classification [6].

ix. GMM classifier: The final classification is performed with a simple GMM classifier. The five feature parameters for each utterance can be viewed as five elements of a feature vector, resulting in a 5-dimensional feature vector for each utterance. A Gaussian mixture model (GMM) with 5-dimension mean and 5x5-dimension covariance can be trained to capture the statistical distribution characteristic of data for each language. The output simply indicates whether the speech utterance is classified as a tonal language or a non-tonal language.

3. PPRLM SYSTEM WITH CONFIDENCE MEASURING

Once the input speech is classified as either a tonal or a non-tonal language, it can be fed into the corresponding PPRLM for the final classification. The PPRLM system for tonal languages is shown in Fig. 3. The PPRLM for non-tonal languages has a similar structure, but it has a different number of phone language models trained by using non-tonal languages.

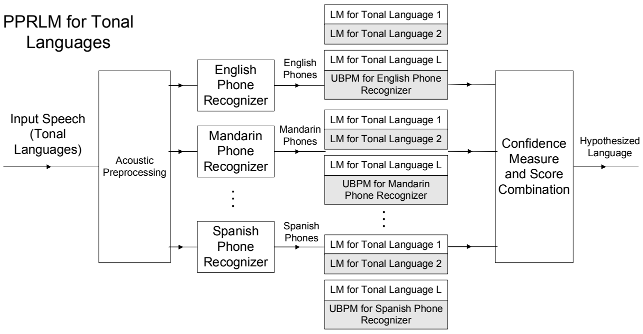

Fig. 3. The Block Diagram of the PPRLM System for Tonal Languages after Pre-classification with UBPM Confidence Measuring

3.1. The Phone Recognizer and Language Model

The phone recognizer maps a speech utterance Ψ into a sequence of phone symbols ψ_P, i.e. $\Psi = \{\psi_1, \psi_2, .. \psi_P\}$, where P denotes the number of phone symbols produced to represent the speech utterance. In this experiment each phone symbol is modelled by a 3-state HMM and each state distribution is modelled by 6 Gaussians. With a given phone recognizer in hand, a N-gram language model is employed to estimate the probability of the occurrence of a particular phone sequence. Considering the amount of training data available in this experiment, we used a tri-gram language model. The perplexity score is used as the output for each testing utterance against each language model, and the language recognition is performed using log-likelihood ratios (LLR).

3.2. Confidence Measure and Score Combination

As mentioned earlier, the number of languages currently spoken in the world is much larger than the number of target languages that current PPRLM systems can handle. This leads us from conventional closed-set forced-choice language identification towards open-set language recognition. Thus this language recognition system should be capable of rejecting non-target languages.

We employ two different tactics for confidence measure and score combination. One is to build a Universal Background Phone Model (UBPM) for each of the phone recognizers; the other is to make the confidence measure not with a background model but by using online garbage models.

LLR-based language recognizers, using UBPM confidence measure with a single phone recognizer can be defined as:

$$LML_i = P(X \mid LM_i) \qquad L_u = P(X \mid UBPM) \tag{7}$$

where LML_i is the likelihood score of the test utterance X for language i's phone language model LM_i, and L_u is the likelihood score of test utterance X for the universal background model UBPM. The recognition score is the log ratio of these two likelihood scores:

$$score_i = \log\{\frac{P(X \mid LM_i)}{P(X \mid UBPM)}\} = \log LML_i - \log L_u \tag{8}$$

For our language recognition system with six phone recognizers, the scores from each language are fused using a linear combination, where k is used to index different phone recognizers:

$$score_i^k = \log LML_i^k - \log L_u^k \quad score_i = \frac{1}{6}\sum_{k=1}^{6} score_i^k \tag{9}$$

For language recognition systems with online garbage models, the UBPM is not employed and the score is defined as the difference between the score of the best hypothesis and the average score of all the language models. For both of these two confidence measures, a threshold score must be determined for either accepting or rejecting the unknown utterance.

4. EXPERIMENTS AND RESULTS

4.1. Corpora Description

The data sources for this experiment were the multi language CALLFRIEND corpus, the OGI-TS corpus and the OGI 22-language corpus [10]. These corpora consist of recorded telephone calls spoken by native speakers of the corresponding languages. Table 1 lists the target and non-target languages used for the evaluation.

There are 6 languages in OGI-TS that are labelled at the fine phone level; these labelled speech utterances were used to train the broad-phone-class recognizer of the pre-classification system and the phone recognizer of PPRLM systems. All data sources for the target languages were from CALLFRIEND and OGI-TS, except for Cantonese which came from the OGI 22-language corpus. For the non-target languages, only Arabic was from CALLFRIEND while the other three languages were from OGI 22-language corpus. The GMM classifier of the pre-classification system was trained with the utterances of target languages from the CALLFRIEND and OGI 22-language corpora. Considering the PPRLM for tonal languages, the UBPM was trained with all the non-target languages and the non-tonal target languages; the UBPM of the PPRLM for the non-tonal languages was trained with all the non-target languages and the tonal target languages. All language models were trained with the corresponding utterances from all three corpora. For evaluation, 30-sec and 10-sec utterances were used. All the evaluation utterances were unseen in training.

Table 1. The sources of the languages used in the experiments

Target Languages	
Cantonese (tonal): OGI 22	English (non-tonal): OGI-TS, CALLFRIEND
Farsi (non-tonal): OGI-TS, CALLFRIEND	French (non-tonal): OGI-TS, CALLFRIEND
German (non-tonal): OGI-TS, CALLFRIEND	Hindi (non-tonal): OGI-TS, CALLFRIEND
Japanese (tonal): OGI-TS, CALLFRIEND	Korean (non-tonal): OGI-TS, CALLFRIEND
Mandarin (tonal): OGI-TS, CALLFRIEND	Spanish (non-tonal): OGI-TS, CALLFRIEND
Tamil (non-tonal): OGI-TS, CALLFRIEND	Vietnamese (tonal): OGI-TS, CALLFRIEND
Non-target Languages	
Arabic (non-tonal): CALLFRIEND	Malay (non-tonal): OGI 22
Russian (non-tonal): OGI 22	Swedish (tonal): OGI 22

4.2. Results

The PPRLM without pre-classification was used as the baseline. Only the four non-target languages were used to train the UBPM for the baseline system. In both the 30-sec and 10-sec evaluations, 120 utterances each from English, Mandarin and Spanish were used, while 60 utterances were used from each of the other 12 languages. In these experiments, a desktop computer with a 3.2GHz single-core CPU and 1GByte of RAM was used, with the front size bus running at 800MHz.

Table 2. Tonal and non-tonal language classification rate (%)

Arabic	*Cantonese*	*English*	*Farsi*
86.7%	**93.3%**	86.7%	86.7%
French	*German*	*Hindi*	*Japanese*
88.3%	83.3%	80%	86.7%
Korean	*Malay*	*Mandarin*	*Russian*
81.7%	88.3%	88.3%	90%
Spanish	*Swedish*	*Tamil*	*Vietnamese*
86.7%	88.3%	83.3%	**95%**

Our results for tonal and non-tonal language pre-classification are given in Table 2. The overall classification rate is 87.1%. Better classification rates are obtained for

Cantonese and Vietnamese. This may be due to the fact that Cantonese has 8 tones and Vietnamese 6 tones, compared with only 4 tones in Mandarin.

In Table 3, we report our recognition rates and processing time given different system configurations. The results demonstrate that pre-classification improves system performance (for both recognition rate and the processing time) in all cases. Also, the language recognition system with UBPM outperforms the language recognition system with online garbage model confidence measures, in most cases. We use the discounting method to fix the probability distribution of the language model, by adjusting low probabilities such as zero probabilities upward, and high probabilities downward. Our previous research indicates that the Witten-Bell discounting method gives the best performance when compared with linear, absolute and Good-Turning discounting [2]. The Witten-Bell discounting method improves the resulting recognition rate slightly. The best recognition accuracy rate of 77.9% is obtained for the 30 sec utterances, by using pre-classification and PPRLM + UBPM + Witten-Bell discounting. This shows a relative improvement of 7.45% compared with an identically configured PPRLM system without pre-classification.

Table 3. Accuracy rate and processing time comparison for different configuration of the language recognition systems (CPU time is the whole system's processing time normalized by the actual length of the corresponding utterance)

	30-sec		10-sec	
Without pre-classification	%Accuracy	CPU time	%Accuracy	CPU time
PPRLM + UBPM	71.1	0.42	45.3	0.42
PPRLM + online garbage model	65.5	0.42	46.4	0.41
PPRLM + UBPM + Witten-Bell Discounting	72.5	0.43	45.5	0.42
	30-sec		10-sec	
With pre-classification	%Accuracy	CPU time	%Accuracy	CPU time
PPRLM + UBPM	73.3	0.39	49.0	0.38
PPRLM + online garbage model	70.4	0.39	47.3	0.38
PPRLM + UBPM + Witten-Bell Discounting	**77.9**	0.39	**49.2**	0.39

5. CONCLUSION

Given that the performance of PPRLM systems decreases as the number of possible languages increases, we have proposed a novel tonal and non-tonal language pre-classification system. Analysis of the speed and level of pitch change is found to be adequate to discriminate between tonal and non-tonal languages. The proposed PPRLM system with tonal and non-tonal language pre-classification, UBPM confidence measure and Witten-Bell discounting is found to be effective and robust, both in terms of recognition rate and processing time. When evaluated with 12 target languages and 4 non-target languages, the novel system can achieve an accuracy rate of 77.9% for 30-sec utterances and 49.2% for 10-sec utterances. Our future work will conduct a more appropriate pre-classification scheme that incorporates other prosodic features such as duration and stress pattern. The new pre-classification system should be capable of classifying unknown languages into a finer subset, such as tonal languages, stress-timed languages, or syllable-timed languages. This will allow for the task of handling increasing amounts of target and non-target languages.

REFERENCES

[1] E. Singer, P.A. Torres-Carrasquillo, T.P. Cleason, W.M. Campbell and D.A. Raynolds, "Acoustic, phonetic and discriminative approaches to automatic language recognition", in *Eurospeech in Geneva,* ISCA, pp. 1345-1348, 2003.

[2] L. Wang, E. Ambikairajah and E. H.C. Choi, "Multi-lingual phoneme recognition and language identification using phonotactic information", in *Proc. ICPR 2006*, vol. 4, pp. 245-248, 2006.

[3] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech", in *IEEE Transactions on Speech and Audio Processing,* vol. 4, no. 1, pp. 31-44, 1996.

[4] C.Y. Lin and H.C. Wang, "Language identification using pitch contour information in the ergodic Markov model", in *Proc. ICASSP 2006*, vol. 1, pp. 193-196, 2006.

[5] B. Ma, D.L. Zhu and R. Tong, "Chinese dialect identification using tone features based on pitch flux", in *Proc. ICASSP 2006*, vol. 1, pp. 1029-1032, 2006

[6] L. Wang, E. Ambikairajah and E. H.C. Choi, "Automatic tonal and non-tonal language classification and language identification using prosodic information", in *Proc. ISCSLP 2006*, pp. 485-496, 2006.

[7] Catford, J.C., *A Practical Introduction to Phonetics*, Oxford University Press, 1988.

[8] Roca, I. and W. Johnson, *A Course in Phonology*, Blackwell Publishing, 1999.

[9] S. Potisuk, M.P. Harper and J.T. Gandour, "Speaker-independent automatic classification of Thai tones in connected speech by analysis-synthesis method", in *Proc. ICASSP 1995*, vol. 1, pp. 632-635, 1995.

[10] Linguistic Data Consortium. http://www.ldc.upenn.edu