



Towards long-range prosodic attribute modeling for language recognition

Raymond W. M. Ng*, Cheung-Chi Leung†, Ville Hautamäki†, Tan Lee*, Bin Ma† and Haizhou Li†‡

*Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

† Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore 138632

‡Department of Computer Science and Statistics, University of Eastern Finland, Finland

*{wmng,tanlee}@ee.cuhk.edu.hk, †{ccleung,vishv,mabin,hli}@i2r.a-star.edu.sg

Abstract

As a high-level feature, prosody may be an effective feature when it is modeled over longer ranges than the typical range of a syllable. This paper is about language recognition with the high-level prosodic attributes. It studies two important issues of long-range modeling, namely the data scarcity handling method, and the model which properly describes prosodic boundary events. Illustrated by NIST language recognition evaluation (LRE) 2009, long-range modeling is shown to bring a 7.2% relative improvement to a prosodic language detector. Score fusion between the long-range prosodic system and a phonotactic system gives an EER of 3.07%. Exploiting boundary N -grams is the main contributing factor to global EER reduction, while different long-range prosodic modeling factors benefit the detection of different languages. Analysis reveals the evidence of language-specific long-range prosodic attributes, which sheds light on robust long-range modeling methods for language recognition.

Index Terms: language recognition, prosody, long-range modeling

1. Introduction

Prosody refers to the rhythmic and intonational properties in speech. The most studied prosodic features include speech fundamental frequency and energy, as well as the duration of speech units. Language recognition, as illustrated in the NIST language recognition evaluation (LRE) tasks [1], is a problem of detecting the presence of a particular language with speech. Prosody is not a conventionally used feature for this problem. However, a number of studies look into the use of prosody in speaker or language recognition [2][3][4][5]. In [5], a comprehensive set of prosodic attributes is shown to provide complementary information to a state-of-the-art language recognition system using the phonotactic approach.

Compared with the typically used cepstral features, prosodic features are suprasegmental and extend over syllables and longer regions [2]. (Pseudo)syllable-based modeling up to trigram and simple modeling in the phrase level are typically employed [3][4][5], but intuitively these higher-level features should be effective when they are modeled in even longer temporal ranges. Prosodic 4-grams and phonetic 5-grams are used in speaker and dialect recognition respectively [6][7]. There has not been a particular study of long-range modeling using prosodic features for language recognition. It would be interesting to see whether higher-order prosodic N -gram features are useful in a language recognition problem.

In the following, a brief introduction to the features and the prosodic attribute model (PAM) is given in Section 2 and 3. Section 4 highlights two considerations for long-range modeling: Skipping models prevent data scarcity; proper ways

are proposed to model boundary prosodic events. Global and language-specific performances using the long-range models are given in Section 5 and 6 respectively.

2. Prosodic attributes

A comprehensive set of prosodic attributes is proved to be effective in language recognition [5]. Three major *types* of prosodic attributes are F0, intensity and duration. Most of the attributes are (pseudo)syllable-based, so feature extraction is done after syllabification. Many attributes are normalized to reduce the undesirable bias to irrelevant factors like speaker variations. Normalization implicitly includes long-range information by capturing the relative feature magnitude with respect to the average measurement over a longer time period.

Some prosodic attributes do not need normalization. For instance, *regression* attributes model the curvature of a pseudosyllabic F0 and intensity contour. Multiple regression attributes are extracted by polynomial regression in different orders. Regression is also done on the contours of two consecutive syllables to model long-range curvature. *Residue* indicates the fluctuations of syllable contours from the phrase curve. It can be considered as the result of normalization against intonation effects.

The prosodic attributes used in this study are summarized in five groups (I) **F0 basic** and (II) **Duration basic** are the groups with different normalization methods. The remaining groups are (III) **F0 regression**, (IV) **Intensity regression** and (V) **F0 residue**. In a previous experiment, the inclusion of some **intensity** attributes causes an error increase [5], thus they are not used in the present study. The total number of prosodic attributes is 45. A concise set of 10 attributes is created by selecting among similar attributes with different normalization methods. These 10 attributes will be used to test for different trigram and higher-order N -gram modeling method. The details of the attributes are shown in Table 1.

3. Long-range modeling of prosody

3.1. Prosodic attribute model

Prosodic attribute model (PAM) is a pseudosyllable-based modeling method for language recognition [5]. PAM is unique for its operation on the prosodic attributes in a parallel and separate manner. In a typical phonetic modeling method, phonetic models cover the whole phonetic space. Whereas in PAM, prosodic units are derived on an attribute-wise basis. N -gram modeling of different prosodic attributes are done separately.

Independence among attributes is a popular assumption for modeling prosody [6]. Compared with an approach when the full prosodic space is modeled altogether, PAM is shown to bring significant reduction in problem dimension and decoding runtime with only a slight increase in error rates [5].

Table 1: Selected prosodic attributes for language recognition

Attribute Group	Measurement [△]	Target contour*
(I) F0 basic	F0 at syllabic nucleus	N/A
	F0 gradient in syllable contour	N/A
(II) Duration basic	Separation between nuclei	N/A
	Voiced-unvoiced ratio	N/A
(III) F0 regression	1 st -order coefficient	1 syllable
	1 st -order coefficient	2 syllables
	2 nd -order coefficient	1 syllable
(IV) Intensity	2 nd -order coefficient	1 syllable
	3 rd -order coefficient	2 syllables
(V) F0 residue	Residual F0 over phrase curve	3 syllables

[△] The term “syllable” refers to pseudosyllable resulted from syllabification

* Target contour is the length of F0 / Intensity contour from which regression results / phrase curve are drawn

3.2. Higher-order N-gram SVM with PAM

Long-range modeling can be realized by either using a longer feature extraction region, or modeling sequential information [2][8]. We focus on the latter approach, typically implemented by N -gram modeling. The separate modeling method of PAM creates parallel attribute groups, in return for fewer models needed for a single attribute. This is beneficial to N -gram modeling. If the bigrams of 39 English phones are modeled, it gives $39^2 = 1521$ bigrams. In the prosodic counterpart, 10 prosodic attributes, each having 6 quantization levels, gives only $10 \times 6^2 = 360$ bigrams.

A popular and effective approach for N -gram modeling in speaker and language recognition is by support vector machine (SVM) [6][9]. For every N -gram of a prosodic attribute, an occurrence *term* records its empirical distribution in a training / test trial. For the $6^2 = 36$ bigrams as described above, a fixed-length vector with 36 terms will be constructed. Recall that PAM models prosodic attributes separately. The fixed-length vectors from different prosodic attributes are concatenated into the final vector, after which SVM trains a language detector. Problem dimension of N -gram SVM for PAM (i.e. number of occurrence *terms*) is affected by the number of prosodic attributes, model size for each attribute and the order and method of N -gram construction.

4. Considerations in higher-order N-grams

4.1. Skipping models

In higher-order N -gram modeling, data scarcity is a concern. It is desirable to avoid zero probability estimates, which happen when an N -gram does not occur in the training data [8]. Assume $w_{n-2}w_{n-1}w_n$ represents the sequence of a particular prosodic attribute across three pseudosyllabic positions (i.e. trigram). In trigram SVM training, a trigram is modeled by one or more occurrence terms. Two configurations are tested:

[FULL] $\overline{C(w_{n-2}w_{n-1}w_n)}$

[SKIPPING] $\overline{C(w_{n-2}w_{n-1}), C(w_{n-2}w_n), C(w_{n-1}w_n)}$

In the [SKIPPING] configuration, a full trigram ($w_{n-2}w_{n-1}w_n$) is broken down into three skipping trigrams ($w_{n-2}w_{n-1}$, $w_{n-2}w_n$ and $w_{n-1}w_n$) to avoid data scarcity. \overline{C} is the occurrence term normalized by the total occurrences of all trigrams in the same training trial. In this study, a training trial is a 30-second speech segment. Extending this idea to skipping 4-gram and skipping 5-gram, $C_2^4 = 6$ and $C_2^5 = 10$ skipping N -grams will be used respectively.

4.2. Boundary prosodic events

In this paper, *boundary* refers to the boundary of a pause-delimited sentence. In long-range modeling, it is more likely that a high-order N -gram touches or spans across a sentence boundary. There are very few related studies on prosody modeling considering boundaries. In [10], consistent reductions of word recognition errors are demonstrated by incorporating sentence boundary information in prosody modeling. For language recognition, pragmatic functions like speaker’s intention and attitude are often expressed near sentence boundaries and may create noise. Examples include tone patterns for interrogations and exclamations. Meanwhile, if language-specific boundary tones exist, modeling sentence boundary will benefit language recognition.

Consider a *boundary bigram*, which is defined as a bigram at sentence initial ($\#_{n-1}w_n$) or at sentence final ($w_{n-1}\#_n$). “#” indicates a sentence boundary. It is inferred from the automatically detected short pauses in speech. By accommodating the skipping model concept, *boundary skipping N-grams*, such as $\#_{n-k}w_n$ and $w_{n-k}\#_n$, can be defined. Three configurations with different treatments to *boundary bigrams* are tested:

[IGNORE] Ignore boundaries: This is the simplest approach among the three. Automatic pause detection is not carried out and there will be no boundary N -grams.

[DELETE] Delete boundary N -grams: The boundary N -grams ($\#_{n-k}w_n$ or $w_{n-k}\#_n$) are dropped before any statistical modeling is done. This is based on the assumption that boundary N -grams mainly carry pragmatic functions unrelated to languages.

[EXPLOIT] Exploit boundary N -grams: A separate pause PAM is constructed for explicit boundary N -gram modeling. For instance, in bigram modeling with 6 models in a prosodic attribute, there are $(6+1)^2 = 49$ bigrams, among which 36 are normal bigrams, 6 are sentence initial bigrams, 6 are sentence final bigrams, and 1 is a pause bigram ($\#_{n-k}\#_n$).

5. Experiments

Results on NIST language recognition evaluation (LRE) 2009 closed-set language detection are reported. The evaluation data includes 10558 30-second utterances of conversational telephone speech and broadcasting speech in 23 languages [1]. Training data include all available data in NIST LRE 1996-2007 and NIST 2009 training data. At least 9 hours of training data is available for each target language. 23 SVM language detectors give the language hypothesis likelihood scores and a number of post-processing steps are performed on the scores. First, a Gaussian backend regulates scores across different language detectors. Second, by making use of a separate development set, scores of multiple language detectors are calibrated following the maximum-a-posteriori (MAP) criterion [11].

The equal error rates (EER) of language detection derived from the calibrated scores of different configurations are compared. Target language dependent detection thresholds will be used. Different configurations are identified by the order of N -grams, by whether skipping N -grams are used and by the methods to model boundary N -grams.

We start with a smaller feature set with 10 selected prosodic attributes (Table 1). Baseline N -gram configuration is the best performing prosodic language recognition setup reported in [5]. The order of N -gram is trigram. It is constructed by skipping-trigrams and boundary N -grams are deleted. Compared with the over 30% EER generally appeared in literature with similar task and features [3][4], this baseline configuration gives an

EER of 20.58%. An optimal long-range modeling method will be found to defeat this competitive baseline system. Finally, the optimal N -gram configuration will be applied to the full set of 45 prosodic attributes, and score fusion with a phonotactic language recognition system will be carried out.

5.1. Full versus Skipping trigrams

The first comparison is between full N -gram models and skipping N -gram models. As the number of full N -grams grows exponentially with N , comparison is confined to trigram models. Each of the 10 prosodic attributes in Table 1 is described by 6 PAMs. This leads to 10 (attributes) $\times 6^3 = 2160$ full trigrams and $10 \times (3 \times 6^2) = 1080$ skipping trigrams. The skipping trigram configuration gives smaller dimensions. Its language recognition EER is 20.58%, compared with 21.41% for the full trigram configuration. Skipping N -grams will be used for constructing higher-order N -grams in subsequent experiments.

5.2. Different modeling to boundary N -grams

Here we compare different boundary N -gram modeling methods. By using skipping N -grams (Sec 4.1), N -grams of different orders are all described by some sets of $6^2 = 36$ skipping N -grams for the [IGNORE] and [DELETE] configurations. In [IGNORE], prosodic events at sentence boundaries are not distinguished from those in other positions. In [DELETE], prosodic events at sentence boundaries are removed. In the [EXPLOIT] configurations, an extra pause PAM is added, giving some sets of $(6 + 1)^2 = 49$ skipping N -grams (Sec 4.2).

The modeling of boundary N -gram has a physical meaning. A skipping N -gram at the boundary ($w_{n-k}\#_n$ or $\#_{n-k}w_n$) explicitly models a pseudosyllable conditioned by its position relative to the sentence boundary (#).

With 10 attributes in Table 1, we compare the language recognition EER conditioned by different boundary modeling methods and different orders of N -grams. Results are shown in Table 2. Language recognition EER is the highest for trigrams with the [IGNORE] configuration. Removing boundary events ([DELETE]) gives a small EER reduction, while the lowest EER is obtained when boundary events are explicitly modeled ([EXPLOIT]). It is observed that exploiting boundary N -grams is the main contributing factor to global reduction of errors.

5-gram modeling exhibits slight but consistent improvements over trigram modeling. Exploiting boundary 5-gram gives an EER of 19.10%, which is a 7.2% relative improvement to the baseline configuration. We expected larger performance gain with long-range modeling, so more analysis will be done in Section 6. Finally, the configuration of exploiting boundary N -grams is extended by using more attributes. EER is 18.46%.

5.3. Fusion with phonotactic system

Scores from the 45-attribute trigram model are fused with the scores from a phonotactic language recognition system which adopts a parallel phone recognition followed by vector-space-model approach. Details of score fusion can be found in [5]. EERs of individual and fused system are shown in Table 3. Compared with [5], the 45-attribute prosodic LRE system with long-range modeling gives lower errors with fewer number of terms. The performance gain of long-range modeling in a prosodic system can be carried forward to the fused system.

6. Language dependent long-range models

Results from Table 2 show that by exploiting boundary N -grams, a 6% relative reduction in global EER can be achieved,

Table 2: Different modeling methods to N -grams at boundary

Boundary N -gram modeling method	Number of attributes / Order of N -gram [†]	EER
IGNORE	10-attribute PAM / trigram	21.27%
DELETE	10-attribute PAM / trigram(baseline)[5]	20.58%
EXPLOIT	10-attribute PAM / trigram	19.40%
	10-attribute PAM / 5-gram	19.10%
EXPLOIT	45-attribute PAM / trigram	18.46%

[†] All N -gram are constructed with skipping N -grams (i.e. sets of distant bigrams)

Table 3: Language recognition performance of a fusion system

	SVM terms in prosodic LRE system	EER		
		Phonotactic	Prosody	Fused
67-attribute PAM[5] ([DELETE],trigram)	7236	3.56%	20.90%	3.18%
45-attribute PAM ([EXPLOIT],trigram)	6027	3.56%	18.46%	3.07%

while 5-gram modeling is marginally better than trigram. The overall performance improvement by long-range modeling is moderate. It is discovered that long-range modeling methods are language and feature dependent. In order to find out which languages and prosodic attributes benefit the most from long-range modeling, we look at the error statistics of particular languages. Because some post-processing steps of the language recognizer involve numerical regulations among the scores from different language detectors, these steps will not be included in this part of analysis.

6.1. Null rejection ratio

We analyze the SVM terms in [EXPLOIT] skipping 5-gram configuration because there are terms corresponding to different long-range modeling factors: 5-gram, boundary trigram and boundary 5-gram. A metric referred to as *null rejection ratio* is used to indicate the contribution of different long-range modeling factors to the detection of a particular target language.

Hundreds of SVM terms represent a prosodic attribute. We fix one term and one target language at a time and test the null hypothesis with one-way analysis of variance (ANOVA). The null hypothesis says that data from the particular SVM term in the target language is the same as that in other languages. We group the SVM terms according to the prosodic attribute they represent, and the modeling factors of trigram, 5-gram, boundary trigram or boundary 5-gram. Then, *null rejection ratio* is derived. It is the proportion of terms, among many sharing the same modeling factors under a prosodic attribute, where the null hypothesis is rejected (at $p < 0.001$). Null rejection ratio takes a value between 0 and 1. The higher a null rejection ratio is, the more robust a modeling factor is. Null rejection ratio of the modeling factor of baseline trigram in every prosodic attribute specific to every target language serves as a reference to indicate the effectiveness of the addition of the other three modeling factors.

6.2. Language dependent analysis

Out of the 23 target languages, 11 target languages have language recognition EER reduced by at least 5% relatively in one or more long-range modeling configurations. Their EER in the baseline trigram configuration and other long-range modeling configurations are included in Table 4. The prosodic attributes which contribute the most in long-range modeling are also found by the inspecting the null rejection ratios.

According to the trends of EER and null rejection ratios, the 11 languages are classified into four groups. Group 1 lan-

Table 4: Analysis of NIST LRE 2009 results with 10 attributes in specific target languages where long-range modeling reduces errors

Language group	Target language	Language recognition EER [†]				Prosodic attribute contributing the most in long-range modeling	Null rejection ratio [‡]			
		trigram* (baseline)	+5-gram*	+Boundary trigram	+Boundary 5-gram		trigram	5-gram	Boundary trigram	Boundary 5-gram
Group 1	Persian	33.33%	31.85%	–	31.53%	3 rd -order intensity regression (2 syllables)	0.66	0.61	0.46	<u>0.74</u>
	Korean	34.74%	32.79%	–	–	3 rd -order intensity regression (2 syllables) F0 gradient in syllable contour	0.34 0.45	0.40 0.45	<u>0.58</u> <u>0.58</u>	0.36 0.52
Group 2	Cantonese	19.26%	–	–	17.99%	1 st -order F0 regression (2 syllables)	0.80	0.82	0.81	0.86
	Vietnamese	7.62%	–	7.27%	7.23%	2 nd -order intensity regression (1 syllable) Residual F0 over phrase curve	0.61 0.90	<u>0.72</u> 0.82	0.67 0.69	<u>0.98</u> <u>1.00</u>
	Hausa	11.31%	–	10.28%	–	2 nd -order intensity regression (1 syllable) 2 nd -order F0 regression (1 syllable)	0.43 0.32	0.40 0.26	<u>0.57</u> <u>0.45</u>	0.43 0.24
Group 3	Amharic	14.32%	–	13.06%	12.56%	2 nd -order intensity regression (1 syllable) 1 st -order F0 regression (1 syllable)	0.61 0.60	0.61 0.46	<u>0.74</u> <u>0.83</u>	<u>0.86</u> 0.60
	Georgian	21.26%	–	20.27%	19.27%	2 nd -order intensity regression (1 syllable) Residual F0 over phrase curve	0.66 0.78	0.69 0.73	<u>0.79</u> <u>0.86</u>	<u>0.86</u> <u>0.88</u>
	Spanish	26.49%	–	24.73%	23.38%	2 nd -order intensity regression (1 syllable) F0 at syllabic nucleus	0.51 0.44	0.51 0.49	0.50 <u>0.63</u>	<u>0.64</u> 0.48
Group 4	Bosnian	23.38%	22.25%	22.25%	21.12%	Voiced-unvoiced ratio	0.20	<u>0.30</u>	<u>0.40</u>	<u>0.33</u>
	Ukrainian	35.05%	–	–	33.03%	2 nd -order intensity regression (1 syllable)	0.44	0.48	<u>0.74</u>	<u>0.62</u>
	Urdu	31.70%	–	–	30.07%	Voiced-unvoiced ratio	0.41	0.44	0.48	0.36
Average (11 languages above)		23.50%	23.05%	22.74%	21.88%	← Long-range modeling reduces EER by 6.9% relatively (without post-processing)				
Average (All 23 languages) [#]		22.20%	22.32%	21.64%	21.34%	← Long-range modeling reduces EER by 3.9% relatively (without post-processing)				

*trigram and +5-gram refer to [DELETE] trigram(baseline) and 5-gram configurations, +Boundary trigram and +Boundary 5-gram refer to the two [EXPLOIT] configurations

[†]EER are based on raw SVM classifier scores without any post-processing. Only EERs which are relatively 5% lower than the baseline is shown.

[‡]Null rejection ratios which is absolutely 0.1 larger than that of the baseline trigram are underlined. [#]For the list of all 23 target languages, please refer to [1]

guages include Persian and Korean. They are the rare languages which benefit from 5-gram modeling. Group 2 are *tonal languages* which benefit from boundary skipping N -grams where N is either 3 or 5. Skipping N -gram is represented by sets of distant bigrams ($\#_{n-k}w_n$ or $w_{n-k}\#_n$), which essentially capture the localized prosodic patterns of the k^{th} syllable (where $k < N$) away from sentence boundaries. The detection of Cantonese benefits from the localized tone patterns far away from sentence boundaries, while Vietnamese and Hausa benefit more from tone patterns towards boundaries. Group 3 includes languages whose detection accuracies continuously improve from boundary trigrams to boundary 5-grams. The combination of these two boundary N -grams suggests some effective modeling on long-range intonation phrases spanning across many syllables. For Group 2 and 3, the boundary N -grams of 2nd-order intensity regression over 1 syllable is important. Group 4 includes languages which have poorly performing baseline attributes. The introduction of new modeling factors just helps.

With the global set of 23 languages, 5-gram modeling reduces errors marginally. Nevertheless, considering only these 11 target languages in Table 4, boundary 5-gram modeling gives lower EER than boundary trigram modeling. This justifies the importance of higher-order N -grams in detecting many, if not all, languages.

7. Conclusion

In this study, long-range prosodic information is modeled for language recognition with the N -gram SVM approach. By exploiting boundary prosodic N -grams, a relative 7.2% reduction of global language recognition error is achieved. This error reduction can be carried forward to a phonotactic-prosodic fusion system. Different long-range modeling methods are effective in different prosodic attributes and different target languages. Our experiments provide some evidence of language-specific long-range prosodic features. Incorporating the knowledge on specific features and languages is expected to help towards more robust language recognition applications.

8. Acknowledgment

This research is partially supported by the General Research Funds (Ref: CUHK 414108) from the Hong Kong Research Grants Council.

9. References

- [1] The 2009 NIST language recognition evaluation results. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/lre/2009/lre09_eval_results
- [2] E. Shriberg, "Higher-level features in speaker recognition," in C. Müller [Ed], *Speaker Classification I*, vol. 4343 of *Lecture Notes in Computer Science/AI*. Springer, Berlin, 2007.
- [3] J.-L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1904-1911, 2007.
- [4] L. Mary, B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Commun.*, vol. 50, no. 10, pp. 782-796, 2008.
- [5] R.W.M. Ng, C.-C. Leung, T. Lee, B. Ma and H. Li, "Prosodic attribute model for spoken language identification," in *Proc. ICASSP*, 2010, pp. 5022-5025.
- [6] E. Shriberg, L. Ferrer, "A Text-constrained prosodic system for speaker verification," in *Proc. Interspeech*, 2007, pp. 1226-1229.
- [7] F.S. Richardson, W.M. Campbell and P.A. Torres-Carrasquillo, "Discriminative N-gram selection for dialect recognition," in *Proc. Interspeech*, 2009, pp. 192-196.
- [8] J. Goodman, "A bit of progress in language modeling," *Comp., Speech and Lang.*, vol. 15, no. 4, 403-434, Oct. 2001.
- [9] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio, Speech, and Lang. Prcs.*, vol. 15, no. 1, pp. 271-284, Jan. 2007.
- [10] A. Stolcke, E. Shriberg, D. Hakkani-Tür and G. Tür, "Modeling the prosody of hidden events for improved word recognition," *Proc. Eurospeech*, 1999, vol. 1, pp. 311-314.
- [11] N. Brümmer, "Application-independent evaluation of speaker detection," in *Comp. Speech and Lang.*, vol. 20, no. 2-3, pp. 230-275, 2006.