VOICED/UNVOICED PATTERN-BASED DURATION MODELING FOR LANGUAGE IDENTIFICATION

Bo Yin[1,2], Eliathamby Ambikairajah[1,2], Fang Chen[2,1]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia
bo.yin@student.unsw.edu.au, ambi@ee.unsw.edu.au, fang.chen@nicta.com.au

ABSTRACT

Most existing duration modeling approaches facilitates phone recognizer and require manually annotated corpus to train the segmentation models, which is usually cost- and time-consuming. In this paper, a novel duration modeling approach is proposed, which does not require phone recognizer/annotated training data, and facilitates fast computation of language identification. In this approach, the segmentation is implemented by using articulatory features like voicing status. A pair of connected unvoiced and voiced segments is considered as the unit, and the duration of each segment is normalized for each utterance and then quantized into 20 discrete ranges. The ranges of units are later considered as symbol sequences and are modeled by n-gram models, to capture the temporal pattern, which is hypothesized to vary in different languages. The experiments based on the NIST LRE 2005 tasks show a relative 19.7% EER improvement by introducing the proposed duration modeling-based system into a fusion system containing two GMM-UBM based acoustic systems using MFCC and pitch+intensity features.

Index Terms – duration modeling, articulatory features, quantization, language identification

1. INTRODUCTION

Recent Language Identification (LID) research has focused on channel normalization, speaker normalization, discriminative modeling and score fusion [1]. Less effort has been put to introduce novel features [2]. While channel/speaker normalization reduces channel mismatch and speaker variation, discriminative modeling techniques expose the boundaries of language classes more accurately, novel features are also critical due to its ability to bring extra information that helps identification.

Duration, as one of the most important components in prosodic rhythm features, carries key information about the semantic construction, prosodic pattern and expressive emotion, which usually vary in different language contexts [3], due to the significantly different way of speech construction. However, duration is not a precisely defined term, because the unit, from which the duration is extracted, must be defined first. Linguistic units such as phonemes, syllables and words are potential choices, and among them phoneme is probably the most widely used one [4, 5].

To extract the durations of phonemes, a phone recognizer has to be deployed to recognize each phone and their boundaries. Durations, then, can be extracted from the timestamps of the boundaries. A modern phone recognizer is usually capable to produce relatively accurate segmentation, but phone-transcribed data has to be prepared for each language to train the language-specific recognizer. A uni-phone recognizer trained from mixed-language corpus could be used instead, but the accuracy is compromised consequently. Finally, the extracted duration of phones are usually modeled by language models as a sequence of symbols.

Phone recognizer-based phone duration modeling would be a well fit approach if there was already a Phone Recognizer followed by Language Model (PRLM) or Paralleled PRLM system constructed and deployed in the target system [5]. However, in a more speed-orientated acoustic system, the extra phone recognizer could be a major set-back when integrating a duration-based system. Furthermore, training the phone recognizers could also be a real pain when adapting the system to a new language.

To take advantage of duration information without the annotation labor, and to get fast acoustic systems, we revisit this topic and propose a novel duration modeling approach. This approach derives duration units from articulatory features such as voicing status, and therefore eliminates the needs of labeling, ensuring the system can be adapted to any language with less effort.

2. UNIT DEFINITION AND DURATION EXTRACTION

Linguistic units commonly used in duration modeling include phonemes, syllables, and words. A word has specific meaning and consists of one or more syllables. A syllable, inherently, contains one or more phonemes, usually in particular patterns such as consonant-vowel-consonant

(CVC), V, CV, CCV, VC, or CVCC [6]. Among all the above linguistic units, a phoneme is usually considered as the smallest structural unit that people can distinguish in spoken languages. By definition, a phoneme does not necessarily stand for a particular sound, but rather an abstraction of a group of similar sounds (called phones) which have similar functions in building language [3]. Therefore, the duration of a phoneme actually is the duration of the corresponding phone in the particular context. The duration of a phone typically ranges from 50 to 200 milliseconds [3], varies in different phonetic, semantic and emotional contexts. Since the phonetic and semantic contexts in different languages are significantly different [3], a hypothesis can be formed that the duration pattern (including the sequence pattern) is language-dependent and therefore can be utilized to discriminate languages. Further experiments are conducted in this research to validate this hypothesis.

While a vowel phone is usually voiced, a consonant phone could either be voiced, unvoiced, or partially voiced [3]. A segment in which all frames are voiced is referred as a voiced segment (V segment) in this paper, and the same scheme applies to an unvoiced segment (U segment). Ideally, voiced and unvoiced phones contain only a single voiced or unvoiced segment, while partially voiced phones may have multiple segments inside. Considering this fact, the voiced/unvoiced segments can be used as the units to represent the fine structure of phones and syllables, and therefore the duration of these units can be used in the duration modeling process. Practically, the error of voicing detection may result in incorrect voiced/unvoiced segmentation, but it is still useful as long as the segments are associated with the linguistic context, which is the case in this research.

The voicing detection algorithm used in this research is based on the ETSI extended front-end for distributed speech recognition [7]. Occasionally, fragments of voiced/unvoiced segments can appear due to voicing detection error or noisy speech. A smoothing algorithm is applied to the detection results that any segment shorter than 20ms is merged to the previous unit, and therefore eliminates the fragments.

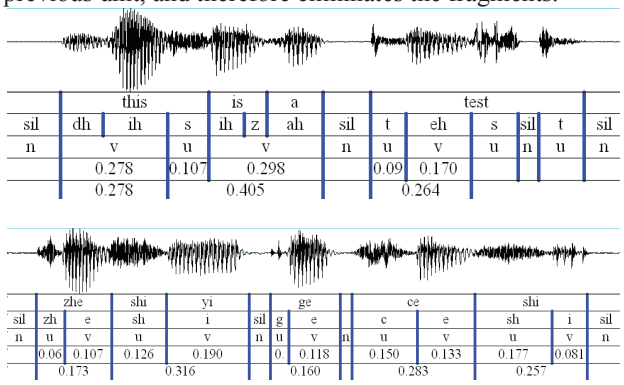

Figure. 1 *An example of voiced/unvoiced pattern in English (upper) and Chinese (lower)*

Fig. 1 shows examples of an English and a Chinese utterance with words (first line), phones (second line) and voicing status (third line) labeled. It is clear from the examples that the boundaries of voiced and unvoiced segments are closely associated with those of phones and/or words. To model the duration, a duration based feature has to be defined first. It has to be decided which voicing unit should be considered as the unit to extract the duration from. The most common sequences of voicing segments include individual voiced (V) segments, unvoiced followed by voiced (UV) segments, and unvoiced followed by voiced and then followed by unvoiced (UVU) segments. Considering the accuracy of automatic segmentation of an unvoiced (U) segment after V segment is usually low, the V and UV segments are focused in this research. Therefore, as a generalized definition, a ***duration unit*** is defined as an optional unvoiced segment followed by a compulsory voiced segment (referred to as UV segment in following sections).

By the definition of the unit, the duration feature can be either based on the UV segment as a whole, or based on the U and V segments separately. For example in the case of the English utterance shown in Fig.1, the duration feature could be

<0.278, 0.405, 0.264>, or

<(0, 0.278), (0.107, 0.298), (0.094, 0.170)>.

The latter definition creates a two-dimensional feature vector from U and V segments within each of the UV units. Similarly the feature of the Chinese utterance could be

<0.173, 0.316, 0.160, 0.283, 0.257>, or

<(0.066, 0.107), (0.126, 0.190), (0.042, 0.118), (0.150, 0.133), (0.177, 0.081)>.

The feature vector containing both U and V durations is supposed to carry more discriminative information, but the higher dimension may result in larger variation in feature space, which therefore hurts the performance. In this research both features are extracted and evaluated. The experiments show that the two-dimensional feature vector incorporating both U and V durations achieved an overall better performance.

3. NORMALIZATION, QUANTIZATION AND SEQUENCE MODELING

The absolute duration values as shown in the bottom two lines of Fig. 1 may vary significantly across utterances regarding to the speaker and context, which makes it difficult to model the duration patterns. To reduce the variation, a simple normalization scheme is introduced:

$$D_{norm} = \frac{D}{D_{avg}} \quad (1)$$

$$(D_{norm\_u},\ D_{norm\_v}) = (\frac{D_u}{D_{avg}}, \frac{D_u}{D_{avg}}) \quad (2)$$

where $\boldsymbol{D}$ is the raw duration of a UV unit, $\boldsymbol{D_{avg}}$ is the averaged duration of the UV units in the current utterance,

D_{norm} is the normalized duration of the UV unit, D_u and D_v are the durations of U and V segments within a UV unit, where D_{norm_u} and D_{norm_v} are the normalized version, respectively. The raw and normalized duration values of the examples in Fig. 1 are shown in Table 1.

To model the temporal variation of durations, a sequence modeling process is introduced. To incorporate the sequence modeling such as n-gram, the normalized duration values need to be quantized into discrete ranges and therefore can be further represented as symbols.

For single-dimensional features, a simple probability based quantization is utilized. For two-dimensional features, the clustering-based quantization technique Vector Quantization (VQ) is investigated in this research, considering its simplicity and previous success in related research [8].

VQ is widely utilized in many lossy compression schemes. The basic idea of VQ is to find the centroids which can be the closest estimation of the majority of feature vectors in feature space. The training process of VQ is based on competitive learning. The distance-sensitivity parameter is introduced in this research for VQ training. The data from all languages are used in training. All duration vectors are quantized into 20 clusters, noted as D1, D2, …, D20. The number of clusters is picked by the optimization process which basically evaluated different cluster numbers including 5, 10, 15, 20, 25, 30 and 35, on the development dataset. For the English example in Fig. 1, the proposed duration feature is a symbol sequence, as shown in Table 1 (Quantized duration).

Table 1. *Examples of normalized and quantized duration features*

Text	This is a test		
Raw UV duration	0.278	0.405	0.264
Normalized UV duration	0.881	1.283	0.836
Raw U/V durations	(0, 0.278)	(0.107, 0.298)	(0.094, 0.170)
Normalized U/V durations	(0, 0.881)	(0.339, 0.944)	(0.298, 0.539)
Quantized duration	D2	D11	D8

While more than one sequence modeling technique is available, n-gram is evaluated in this research due to its success in existing phone and phone duration sequence modeling approaches [5].

To model the symbol sequence such as <D2, D11, D8> produced by the quantization stage, both bi-gram and tri-gram are implemented and evaluated. Tri-gram achieved superior performance in a larger database while bi-gram performed better in a smaller database.

4. EXPERIMENTS

The proposed duration-based system as shown in Fig.2 is first evaluated on the the OGI database individually, with various parameters. The best performed configuration is then evaluated on the CallFriend database according to the NIST LRE 2005 guidelines. A fusion-based system which incorporates the proposed duration-based system and other acoustic systems is developed and evaluated to investigate if the proposed system contributes complementary information to the task.

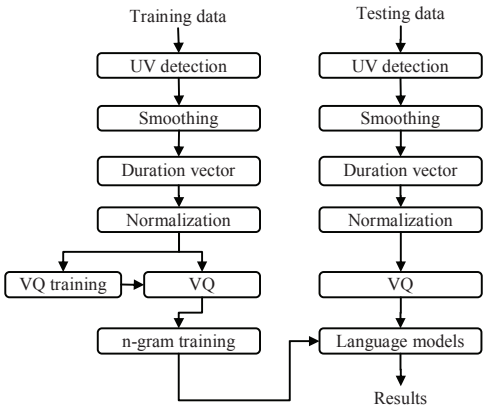

Figure 2. *The proposed duration modeling based system*

The OGI telephony speech database is a multi-language, multi-speaker corpus, composed of a minimum of 90 calls (approx. 2 minutes each, different speakers for different calls) in each of the 10 languages. 50 calls were used as the training set, 20 as the development set and the remaining 20 as the evaluation set. Two different duration features are evaluated – the one-dimensional for the whole UV unit and two-dimensional for separate U and V segments - as discussed in section 2. Bi-gram and tri-gram are both investigated, with each of the features. The evaluation results are shown in Table 2.

Table 2. *Accuracy of the proposed duration-based system with different features and sequence models*

SYSTEM	ACCURACY
UV duration based 1-dimensional feature bi-gram	72.3%
U/V durations based 2-dimensional feature bi-gram	**81.1%**
UV duration based 1-dimensional feature tri-gram	64.2%
U/V durations based 2-dimensional feature tri-gram	69.7%

As shown in Table 2, the two-dimensional feature based on separate U and V segment durations performed remarkably better than the one-dimensional feature based on the whole UV unit. It can be explained that the separate U and V segments contain more phonetic or linguistic fine-structure information which contributes to discriminate languages. Additionally, bi-gram models achieved significantly higher accuracy than tri-gram, partly because the training data is insufficient for tri-gram training.

The best performed duration based systems are evaluated on CallFriend database, then, according to the NIST LRE 2005 guidelines, to cross-validate the effectiveness of the proposed duration modeling approach. The CallFriend database contains 60 half-hour telephone conversations. 7 languages from the database are used in this evaluation.

Besides the evaluations as an individual system, the duration based system is also integrated into a fusion based system which contains two GMM-UBM based acoustic systems incorporating MFCC and pitch+intensity features, Shifted Delta Coefficients (SDC), and Feature Warping, which are detailed in [9, 10]. Fig. 3 illustrates the system.

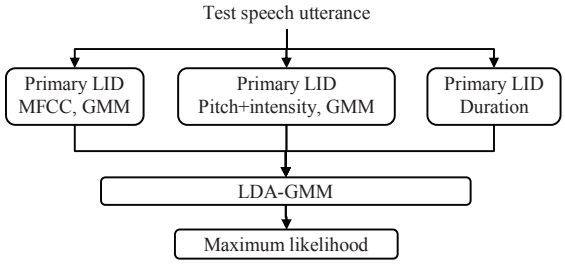

Figure 3. *The fusion based system integrating the proposed system*

Different systems were evaluated on the NIST LRE 2005 task (30s only, primary condition, 7 languages). The performances are reported in Table 3.

Table 3. *Equal Error Rates (EER) of various systems on NIST LRE 2005 task (30s, primary condition)*

SYSTEM	EER%
Duration based system, bi-gram, 2-D feature	17.3
Duration based system, tri-gram, 2-D feature	14.8
Fusion based system, tri-gram, 2-D feature MFCC, Pitch+Intensity	7.1
Fusion based system, tri-gram, 2-D feature MFCC, Pitch+Intensity, Duration	**5.7**

Interestingly, opposite to the experiments on the OGI database, the tri-gram based duration system outperformed the bi-gram based system on the NIST LRE 2005 task. This is potentially because the larger-size CallFriend database provides better coverage of samples for tri-gram training. Consequently, the tri-gram system was deployed in the final fusion system. Apart from the reasonable performance achieved by the individual duration based systems, fusing the proposed duration system into existing acoustic systems introduced a relative 19.7% EER reduction.

5. DISCUSSION

In this paper, a novel and systematic duration modeling approach is proposed and evaluated, which derives the duration features from voicing status, therefore eliminates the manual annotation required by a phone recognizer in conventional duration modeling methods. Considering this advantage, the proposed voiced/unvoiced pattern based duration modeling system is well fitted when combined with other acoustic systems to produce a fast system which is also easy to adapt to new languages. As a limitation, the proposed approach relies on an accurate voiced/unvoiced detection. The detection errors may degrade the performance if it is not consistent. Therefore, the phone-based duration modeling might be more appropriate if phone recognizers already exist in the system.

In the future, alternative quantization and sequence modeling techniques need to be further explored. The application of the proposed duration modeling approach in other speech classification tasks should also be investigated.

6. REFERENCES

[1] D. Reynolds, "Speaker and Language Recognition - A Guided Safari," in *Odyssey Workshop*, South Africa, 2008.

[2] B. Yin, T. Thiruvaran, E. Ambikairajah, and F. Chen, "Introducing a FM based Feature to Hierarchical Language Identification," in *InterSpeech*, Brisbane, 2008.

[3] C. John, Y. Collin, and F. Janet, *Introduction to Phonetics and Phonology*: Oxford: Blackwell, 2007.

[4] D. Povey, "Phone duration modeling for LVCSR," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 2004, pp. I-829-32 vol.1.

[5] M. A. Zissman, "Language identification using phoneme recognition and phonotactic language modeling," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, 1995, pp. 3503-3506 vol.5.

[6] G. N. Clements and S. J. Keyser, "CV phonology: A generative theory of the syllable," *Linguistic inquiry monographs* vol. 9, 1983.

[7] T. Ramabadran, A. Sorin, M. McLaughlin, D. Chazan, D. Pearce, and R. Hoory, "The ETSI extended distributed speech recognition (DSR) standards: server-side speech reconstruction," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 2004, pp. I-53-6 vol.1.

[8] C. Sin-Horng, L. Wen-Hsing, and W. Yih-Ru, "A new duration modeling approach for Mandarin speech," *Speech and Audio Processing, IEEE Transactions on,* vol. 11, pp. 308-320, 2003.

[9] B. Yin, E. Ambikairajah, and F. Chen, "Hierarchical Language Identification based on Automatic Language Clustering," in *InterSpeech - EuroSpeech*, Antwerp, Belgium, 2007.

[10] B. Yin, E. Ambikairajah, and F. Chen, "Improvements on hierarchical language identification based on automatic language clustering," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4241-4244.