Interspeech 2018
2-6 September 2018, Hyderabad

Sub-band Envelope Features using Frequency Domain Linear Prediction for Short Duration Language Identification

Sarith Fernando[1,2], Vidhyasaharan Sethu[1], Eliathamby Ambikairajah[1,2]

[1]School of Electrical Engineering and Telecommunications, UNSW Sydney
[2]DATA61, CSIRO, Sydney, Australia
sarith.fernando@unsw.edu.au

Abstract

Mismatch between training and testing utterances can significantly degrade the performance of language identification (LID) systems, especially in the case of short duration utterances. This work explores the hypothesis that long-term trends are less affected by this mismatch compared to short-term features. In particular, it proposes the use of features based on temporal envelopes within sub-bands. In this work, the temporal envelopes are obtained using linear prediction in the frequency domain. These envelopes are then transformed into cepstral features. The proposed features are then used as a front-end to a bidirectional long short term memory recurrent neural network to identify languages. Experimental evaluations on the AP17-OLR dataset under different conditions indicate that the proposed features exhibit substantially greater robustness under different noise and mismatch conditions, compared to baseline features. Specifically, the proposed features outperform state-of-the-art bottleneck features and show a relative improvement of 38.4% averaged across the test set.

Index Terms: Language Identification, Frequency domain linear prediction, Bottleneck features, Temporal envelope features, Bidirectional modelling

1. Introduction

The use of phonotactic features for language identification (LID) is the most promising approach to building LID systems to date [1-3]. Advances in deep learning methods to extract phonotactic features make the process much easier due to the availability of large data sets and their modeling capabilities. Over all, deep bottleneck features (BNF) [4] have achieved significant performance gains for LID. It is hard to find a recent system that does not use a BNF based front-end as a feature extractor for LID tasks [5-7]. The variations of speakers, background noise, speech content and channel effects make the LID task more challenging. Hence, it is important to explore language characteristics and information of speech statistics rather than the speech itself. Statistics of the phonemic constraints and language information are captured by either BNF or the phonotactic representation output from a phone recognizer (e.g. PLLR features) [1]. However, the performance of a LID system relies heavily on the effectiveness of the phone recognizer and also consumes significant amounts of time.

On the other hand, the spectral distribution of each language is mainly captured by acoustic features. Extracting these acoustic features is an efficient process and there is no requirement for linguistic information. The most popular acoustic features for LID are shifted delta coefficients and eigen features [3], estimated from traditional Mel-frequency cepstral coefficients [8] and perceptual linear prediction features [9]. Typically, when using phonotactic or acoustic features, the commonly used system is total variability modelling (i-vectors) followed by a GMM-UBM back-end. Despite all the recent advances in LID, short duration utterances still suffer significant performance degradation due to various mismatch conditions [10, 11]. Interestingly, backend modelling by deep neural network (DNN) based approaches showed promising gains compared to well-known UBM i-vector approaches [12, 13]. Even though there are different DNN based approaches to short duration LID tasks, bidirectional long short term memory (BLSTM) recurrent neural network based LID systems are able to achieve significant performance gains by capturing robust sequential information from the given input features [14].

In this paper our primary work is to introduce novel features that have higher descriptive and discriminative capabilities for short duration LID tasks. We develop a LID system using features extracted from sub-band temporal envelopes, followed by end-to-end bidirectional modelling.

2. Features for LID

Speech analysis typically relies on spectral content estimated from short windows of about 10-20ms, and produces frame level short term spectral features [8]. These acoustic features mostly carry information related to formants. Consequently, the dynamics of the speech signal are captured by the derivatives of these features [3]. However, these feature extraction techniques may be highly vulnerable in the presence of convolutive and additive nose, i.e. mismatches between training and testing data. Although the BNF proposed in [4], perform well for LID, these features are still based on short term spectral features. Therefore, use of these features for so-called DNN based approaches, including BLSTM LID systems, causes significantly degraded performance due to mismatch conditions [6]. The two most prominent mismatch conditions are channel and duration mismatch. Duration mismatch occurs when the system is trained on fixed length utterances and tested on different length utterances. DNNs are also generally trained on sequence lengths on the order of tens of frames, and thus have poor performance when evaluated on utterances that are a few seconds long [15]. Channel mismatch occurs when training and testing data comes from two different channels.

An alternative way to analyze speech signals is to look for long term temporal features which may help reduce the mismatch in duration while training and testing sequence based systems like BLSTMs. In [16], speech envelope based frequency domain linear prediction (FDLP) features were

10.21437/Interspeech.2018-1805

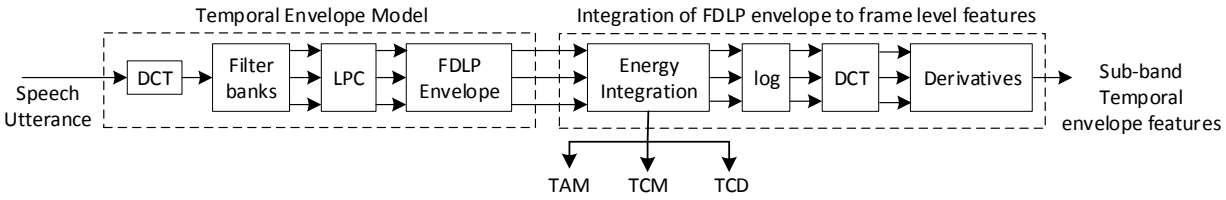

Figure 1: *Proposed feature extraction schematic with three types of energy integration methods to compute envelope features.*

proposed for robust phoneme recognition and were shown to have higher accuracy rates. Borrowing the concept of FDLP, we investigate temporal envelope based features for bidirectional modelling of short duration language identification task in this paper.

2.1. Frequency domain linear prediction

Spectral domain linear prediction was first proposed in [17], and later evolved into the so-called frequency domain linear prediction (FDLP) [16] that represents a smoother version of the instantaneous energy of the speech signal. The FDLP envelope captures the structure of speech, e.g. the amplitude modulation component, whereas the residual error of the FDLP envelope characterizes fine variations of the signal, like frequency modulated components. We believe that this approximated smoother envelope might able to mitigate mismatch, leading to better feature modelling ability for LID tasks.

2.2. Proposed temporal envelope features for LID

Figure 1 shows a schematic diagram for the implementation of FDLP. First, the discrete cosine transform (DCT) is applied to the input speech signal $s[n]$ with N samples, and transformed into the frequency domain equivalent $C[k]$, where $k = 0,1 \dots N-1$, as

$$C[k] = \sqrt{\frac{2-\delta_k}{N}\sum_{n=0}^{N-1} s[n] cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right]} \quad (1)$$

where δ_k is 1 when $k = 0$ and 0 otherwise. The i^{th} sub-band DCT component is then yielded by multiplying the full-band DCT with the i^{th} mel filter $H^i[k]$ as

$$Y^i[k] = C[k] \cdot H^i[k] \quad (2)$$

The output from each filter bank Y^i is used to calculate linear prediction coefficients a_r^i using the autocorrelation method [18]. Finally, the temporal envelope of the speech signal from each filter is calculated by taking the 'frequency' response of the linear predictor,

$$\hat{s}^i[g] = \left|\frac{1}{1+\sum_{r=1}^{b} a_r^i e^{-j2\pi g}}\right|^2 \quad (3)$$

where b is the number of poles and the response is evaluated at points $g = [0, 1, \dots, G-1]$. The specific energy maxima in the time domain signal can be directly correlated with the resulting individual poles. By fixing a specific number of poles in a given interval, we tend to calculate only a fixed number of distinct energy peaks. Unlike any other peak detection method, the above method has the ability to approximate these dominant peaks well and remove finer-scale details, which is beneficial for LID tasks.

To reduce the number of samples in the estimated signal $\hat{s}^i$ from band i, we introduce three types of energy integration methods to estimate compact feature representations: temporal average magnitude (TAM), temporal centroid magnitude (TCM), and temporal centroid distance (TCD).

TAM: First, the FDLP envelope sequence $\hat{s}^i[g]$ is multiplied by a hamming window $w_h[z]$, then averaged across the window as

$$TAM^i[p] = \frac{1}{L}\sum_{z=1}^{L} \hat{s}^i[pM-z] \cdot w_h[z] \quad (4)$$

where the window length L is smaller than the sequence length N, p is frame index, and M is the separation between frames.

TCM: The TCM is the weighted average (centroid) magnitude for a given time frame, given as

$$TCM^i[p] = \frac{\sum_{z=1}^{L} \hat{s}^i[pM-z] \cdot r^i[pM-z]}{\sum_{z=1}^{L} r^i[pM-z]} \quad (5)$$

where, $r^i[g]$ is a suitable weighting term. The TCM captures the first order energy distribution at each frame.

TCD: In order to obtain an abstract representation of the centroid magnitude variation with respect to the centroid of each window as,

$$TCD^i[t] = \left|\frac{\sum_{z=1}^{L} \hat{s}^i[pM-z] \cdot r^i[pM-z]}{\sum_{z=1}^{L} \hat{s}^i[pM-z]} - \frac{\sum_{z=1}^{L} r^i[pM-z]}{L}\right|^{-1} \quad (6)$$

This can detect the approximate location of formants in each frame.

3. Feature extraction and experimental setup

The complete experimental setup is shown in Figure 2. In this paper our main goal is to evaluate the proposed features for short duration language identification when there is mismatch between training and testing data. For this task we use the AP17-OLR dataset [15] which consists of 10 different languages and was specifically developed for short duration language identification tasks. The database also has training and test speech from two diverse sources. Three languages (Japanese, Russian and Korean) are recorded in two different environmental conditions (designated 'mismatched' in this work), whereas all other languages have only one condition ('matched'). The duration of training data for each language is about 10 hours of speech sampled at 16 kHz. The proposed system was tested on three duration conditions, namely, 1sec,

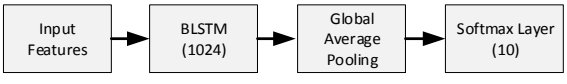

Figure 2: *Block schematic of an end-to-end BLSTM architecture of LID system.*

3sec and 'all' duration subsets of the development set. These subsets contain 17964, 16404 and 17964 utterances respectively. Given the short duration of each test utterance, no voice activity detection was employed.

The backend is a simple BLSTM system consisting of a single hidden layer with 1024 units, and 10 classes as the outputs. After the BLSTM layer, feature level averaging is conducted before feeding into a softmax layer. Training is carried out using truncated back propagation through time with 1s duration utterances. The proposed temporal envelope features (Section 3.2) are compared with baseline features BNF and MFCC (Section 3.1). These baseline techniques are chosen as they are commonly deployed in LID tasks.

3.1. Bottleneck and MFCC feature extraction

The BNF extraction process is based on a phonetic model, time-delay neural network (TDNN) [19] and was trained using the THCHS30 database. The raw input features to the TDNN are 40-dimensional Mel-filter bank coefficients, with a symmetric 4-frame window for the TDNN. The TDNN has 6 hidden layers, and the activation function is p-norm. The number of units of each TDNN layer is set to be 2048, except for the last hidden layer, which has only 256 units.

The static MFCC features involved 47 Mel-filter banks using a 25-ms window and 10-ms frame-shift. Then 13-dimensional Mel frequency cepstral coefficients (MFCCs) were calculated for each frame [20]. These static features were augmented by their first and second order derivatives, resulting in 39-dimensional feature vectors.

3.2. Temporal envelope feature Extraction

The proposed features are extracted using 47 mel filters implemented in the DCT domain over a one second window ($N = 16000$ given a sampling rate of 16kHz), and employing a 160th order linear prediction model in each sub-band. The 'frequency response' of linear predictor is evaluated at 400 sample points ($G = 400$ in eqn. (3)). Frame level temporal envelope features are estimated by the three energy compression methods of TAM, TCM and TCD (Section 2.2), choosing a 25ms window (L=10) with 10ms frame shift (M=4). This will result in the same number of frames as the baseline BNF and MFCC feature estimation methods.

The weighting term $r^i[g]$, used for TCM and TCD computation is chosen as,

$$r^i[g] = f_l^i + \frac{f_u^i - f_l^i}{G}\, g \qquad (7)$$

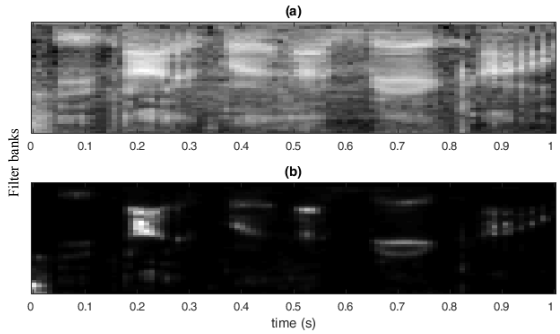

Figure 3: *Comparison of (a) an MFCC spectrogram, and (b) a TAM spectrogram representation for a 1s duration of Cantonese Chinese recording.*

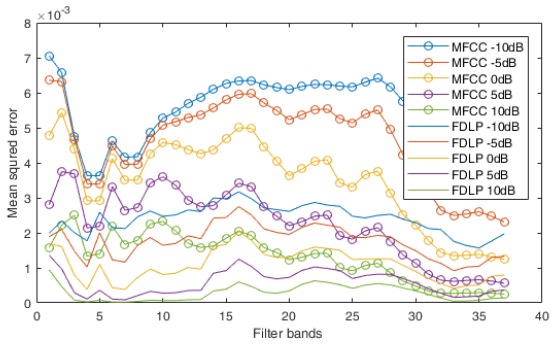

Figure 4: *Comparison of MFCC and TAM features in different babble noise conditions (-10:5:10dB SNR).*

where f_u^i and f_l^i are the highest and lowest frequencies of the i^{th} sub-band and G is the number samples in $\hat{s}^i[g]$.

Finally, similar to MFCCs, cepstral coefficients were calculated for the extracted temporal envelope features and augmented with their derivatives.

4. Analyzing temporal envelope features for LID

In this section, we compare the differences between the proposed TAM features and the baseline BNF and MFCC features. Figure 3 compares a spectrogram obtained with a mel filterbank, to a spectrogram representation of the sub-band envelopes estimated by the proposed method for a Cantonese speech segment of duration 1sec. It can be seen that the proposed TAM spectrogram is less noisy compared to the mel spectrogram while retaining the spectro-temporal energy structure in speech.

4.1. Robustness in channel noise

Typically, a noisy speech signal can be modelled with the addition of convolutional channel noise and additive background noise. Here, there is a need for robust features that suppress these type of noise distortions. Figure 4 compares the mean squared error between features in clean conditions and those extracted under several noisy conditions (babble noise of varying SNR, babble noise was chosen since it has similar spectral characteristics to the speech signal and hard to distinguish from it) for both MFCCs and the proposed TAM features, where the features are estimated from 10% of the training data from each language. It is evident from Figure 4 that the TAM envelope provides greater robustness to noise.

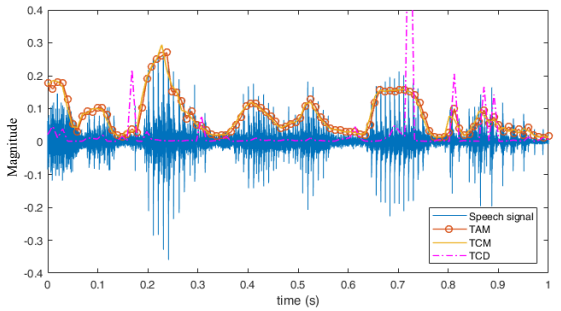

Figure 5: *Comparison of TAM, TCM and TCD for a 1s duration of Cantonese Chinese recording.*

This phenomenon is directly applicable when training and test data comes from two different sources, i.e. mismatched conditions.

In order to study the effect of the finer spectral resolution in the proposed feature extraction technique, Figure 5 shows the analysis of the dynamics of TAM, TCM and TCD features. The variation with respect to time of each feature has a similar pattern. However, while TAM and TCM sit nearly on top of each other, TCD shows quite different variations.

5. Experiments and results

Table 1 compares the performance of the baseline BNF and TAM features in terms of Cavg, which is the primary evaluation measure for the AP17-OLR dataset [15]. It is clear that TAM has significant improvement of 25.62% relative to the baseline BNF features for 1s duration utterances. Note that this improvement is much greater (34.6%) in 'mismatch' condition languages (Japanese, Russian, and Korean) compared to 'matched' languages (17.66%). Therefore, it is evident that TAM features are more robust to mismatch conditions.

Table 1: Performance of the proposed TAM features compared to BNF for AP17-OLR 1s duration for matched and mismatched conditions.

Condition		Cavg		Improvement [%]
		BNF	TAM	
1	Matched	0.0923	0.0760	17.66
2	Mismatched	0.1497	0.0979	**34.60**
Overall		0.1214	**0.0903**	25.62

To investigate the performance of the proposed feature extraction methods, we evaluated the system performance for different durations of utterances. The LID results for the baseline systems and the proposed envelope features are shown in Table 2. All features in Table 2 were tested using the BLSTM system (Section 3) except for BNF_LSTM system, which uses the LSTM system in [15] (included for fair comparison with other systems to date). The results suggest that the proposed envelope features improve performance for all three duration conditions, providing evidence for the suppression of mismatch of duration in training and testing conditions. Even though the BLSTM system was trained on 1s duration utterances, the highest gain when comparing BNF and TCM (55.5%) comes from the 3s duration test condition, decreasing Cavg from 0.0668 to 0.0297. The TAM results show an average 38.4% relative improvement with respect to the BNF, and suggest that the modeling of high-energy regions in time-frequency domain is beneficial in mismatch conditions.

Table 2: System performance in different feature extraction methods

Feature	Performance (Cavg)			
	1s	3s	all	Avg
MFCC	0.1277	0.1132	0.1021	0.1143
BNF_LSTM [15]	0.1153	0.0727	0.0689	0.0856
BNF	0.1214	0.0668	0.0589	0.0824
TAM	**0.0903**	0.0319	0.0303	**0.0508**
TCM	0.0943	**0.0297**	**0.0294**	0.0511
TCD	0.1398	0.1297	0.1236	0.1310
TAM+TCD	0.0649	0.0226	0.0165	0.0347
BNF+TAM	0.0520	0.0162	0.0156	0.0279
BNF+TAM+TCD	**0.0489**	**0.0159**	**0.0130**	**0.0259**

Moreover, the TCM features achieved similar performance to the TAM features. On the other hand, the TCD features fail to outperform any other features independently, so score fusion was conducted using the Focal toolkit [21] to compare feature interdependency further. The fusion of TCD and TAM features shows the availability of the complementary information in TCD, showing its benefit as an additional feature set for LID tasks. Moreover, BNF and TAM features are also complementary. The best fusion system performance gain was by BNF, TAM and TCD systems together, with an average relative improvement of 68.57% over the baseline BNF.

6. Conclusions

In this paper, we have proposed envelope based frame level, frequency domain linear prediction (FDLP) features (TAM, TCM and TCD) for short duration language identification tasks. Estimates of temporal envelopes within sub-bands of speech are initially obtained and the features are extracted from these envelopes. Various experiments are performed with AP17-OLR data, where the proposed features provide significant improvements over state-of-the-art BNF features. Further, we showed the robustness of features compared to existing spectral features, specifically MFCCs. Compared to the existing bottleneck features the results are promising and encourage further investigation of the FDLP domain for LID tasks.

7. References

[1] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/ ,Brno, Czech Republic, 2008.

[2] S. Fernando, V. Sethu, and E. Ambikairajah, "A Feature Normalisation Technique for PLLR Based Language Identification Systems," in *INTERSPEECH*, 2016, pp. 2925-2929.

[3] S. Fernando, V. Sethu, and E. Ambikairajah, "Eigenfeatures: An alternative to Shifted Delta Coefficients for Language Identification," presented at the SST2016, Parramatta, Australia, 2016.

[4] F. Richardson, D. Reynolds, and N. Dehak, "A Unified Deep Neural Network for Speaker and Language Recognition," *arXiv preprint arXiv:1504.00923*, 2015.

[5] B. Jiang, Y. Song, S. Wei, J.-H. Liu, I. V. McLoughlin, and L.-R. Dai, "Deep bottleneck features for spoken language identification," *PloS one*, vol. 9, p. e100795, 2014.

[6] S. Fernando, V. Sethu, and E. Ambikairajah. (2018, Hidden variability subspace learning for adaptation of deep neural networks. *Electronics Letters 54(3)*, 173-175. Available: http://digital-library.theiet.org/content/journals/10.1049/el.2017.4027

[7] S. Fernando, V. Sethu, and E. Ambikairajah, "Factorized Hidden Variability Learning for Adaptation of Short Duration Language Identification Models," presented at the ICASSP, Calgary, Alberta, Canada, 2018.

[8] E. M. Mohammed, M. S. Sayed, A. M. Moselhy, and A. A. Abdelnaiem, "LPC and MFCC performance evaluation with artificial neural network for spoken language identification," 2013.

[9] B. Yin, E. Ambikairajah, and F. Chen, "Combining cepstral and prosodic features in language identification," in *Pattern*

Recognition, 2006. ICPR 2006. 18th International Conference on, 2006, pp. 254-257.
[10] M.-G. Wang, Y. Song, B. Jiang, L.-R. Dai, and I. McLoughlin, "Exemplar based language recognition method for short-duration speech segments," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7354-7358.
[11] R. Travadi, M. V. Segbroeck, and S. S. Narayanan, "Modified-prior i-vector estimation for language identification of short duration utterances," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
[12] J. Gonzalez-Dominguez, I. Lopez-Moreno, P. J. Moreno, and J. Gonzalez-Rodriguez, "Frame-by-frame language identification in short utterances using deep neural networks," *Neural Networks*, vol. 64, pp. 49-58, 2015.
[13] A. Lozano-Diez, R. Zazo Candil, J. González Domínguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "An end-to-end approach to language identification in short utterances using convolutional neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.
[14] S. Fernando, V. Sethu, E. Ambikairajah, and J. Epps, "Bidirectional Modelling for Short Duration Language Identification," presented at the Interspeech 2017, Sweden, 2017.
[15] Zhiyuan Tang, Dong Wang, Yixiang Chen, and Q. Chen, "AP17-OLR Challenge: Data, Plan, and Baseline," *arXiv:1706.09742*, 2017.
[16] S. Ganapathy, S. Thomas, and H. Hermansky, "Temporal envelope compensation for robust phoneme recognition using modulation spectrum," *The Journal of the Acoustical Society of America*, vol. 128, pp. 3769-3780, 2010.
[17] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *The Journal of the Acoustical Society of America*, vol. 105, pp. 1912-1924, 1999.
[18] L. B. Jackson, *Digital Filters and Signal Processing*: Springer US, 1996.
[19] K. J. Lang, A. H. Waibel, and G. E. Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural networks*, vol. 3, pp. 23-43, 1990.
[20] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: a tutorial," *Circuits and Systems Magazine, IEEE*, vol. 11, pp. 82-108, 2011.
[21] "FoCal, Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers http://sites.google.com/site/nikobrummer/focal," 2008.