ON THE IMPORTANCE OF ANALYTIC PHASE OF SPEECH SIGNALS IN SPOKEN LANGUAGE RECOGNITION

*Karthika Vijayan, Haizhou Li**

Dept. of Electrical and Computer Engineering
National University of Singapore
{vijayan.karthika, haizhou.li}@nus.edu.sg

Hanwu Sun, Kong Aik Lee

Human Language Technology Dept.
Institute for Infocomm Research
A*STAR, Singapore
{hwsun, kalee}@i2r.a-star.edu.sg

ABSTRACT

In this paper, we study the role of long-time analytic phase of speech signals in spoken language recognition (SLR) and employ a set of features termed as instantaneous frequency cepstral coefficients (IFCC). We extract IFCC from long-time analytic phase, in an effort to capture long range acoustic features from speech signals. These features are used in combination with the traditional shifted delta cepstral coefficients (SDCC) for SLR. As the SDCC are extracted from spectral magnitude and IFCC are from analytic phase, they characterize long-time information of speech in different ways. The experiments conducted with NIST LRE 2017 task reveals the complementary effects of IFCC features to SDCC and deep bottleneck (DBN) features. The fusion of IFCC with SDCC/DBN features delivered relative improvements of 23.23% and 16.78% in average equal error rate over the SDCC and DBN features, respectively, indicating the benefits of information from analytic phase in SLR.

***Index Terms*—** Spoken language recognition, Long-time features, Analytic phase, Instantaneous frequency, Fusion.

1. INTRODUCTION

The task of spoken language recognition (SLR) refers to identification or verification of the language identity of a speech segment [1]. An efficient SLR system is essential for applications such as, multilingual speech recognition, automatic translation of speech recordings and audio information retrieval. Generally, an SLR system relies on phonetic and phonotactic information in speech. The phonetic information can be linked to acoustic events generated by human vocal tract system, and hence to segmental short-time features of speech. But, the phonotactic information in speech are suprasegmental, as they are linked to sequence of phones/syllables and rules/syntax governing combinations of phones forming words, phrases, etc. There have been a key interest in acoustic features that characterize long-time characteristics of speech signals [1].

A prominent method to capture phonotactic information is the parallel phone recognizer followed by language model (PPRLM), in which multiple language-dependent phone recognizers were trained using short-time features followed by an n-gram language model learning likely sequences of phones [2, 3]. Approaches to improve effectiveness of PPRLM includes usage of phone lattices [4], selection of discriminative phone-subsequences [5] and phone log likelihood ratios (PLLR) as features [6, 7]. The disadvantage of using n-gram language models is that the number of n-grams increases exponentially with the attempt to capture longer phonotactic context by increasing the length of phone sequences [1].

The shifted delta cepstral coefficients (SDCC) are features incorporating long range dynamic characteristics in speech signals. The SDCC features for a particular short-time frame consist of delta values between mel frequency cepstral coefficients (MFCC) from multiple neighboring frames [8, 9]. The SDCC were widely used as features for SLR, employing various classifiers like, Gaussian mixture models (GMM) [3, 8, 10–12], support vector machines (SVM) [3, 13], deep neural networks (DNN) [14], etc. Another front-end for SLR was developed by training a deep bottleneck (DBN) network using short-time features [15–18]. The DBN features learn long-time information by capturing inter-frame relationships from multiple short-time frames, which will add the requirement for additional training and associated computational complexity to SLR system. And, the SDCC and DBN features represent long-time information from short-time spectral magnitude of speech signals.

The prosody of speech also contributes to language-specific cues for SLR. Major prosodic elements in speech like, loudness, duration, rhythm & intonation, pitch, etc. were investigated for feature extraction in SLR [19–22]. Another set of features from short-time phase spectrum (modified group delay features) was also used for SLR [23]. These features, upon combination with SDCC features, had delivered promising improvement to SLR performance. Unfortunately, all the phonotactic and prosody features are extracted based on short-time spectral analysis, that inherently suffers from issues such as frequency leakage and phase wrapping.

In this paper, we study the importance of *analytic phase* in characterizing language traits and perform feature extraction through long-time analysis of speech signals. The *instantaneous frequency* (IF) is computed as representative of long-time analytic phase, without being affected by the phase wrapping problem. The IF cepstral coefficients (IFCC) are extracted by segmenting IF into short time frames. The SDCC and IFCC features are extracted from spectral magnitude and analytic phase, respectively. Both these features contain long-time information from two different domains of transformation of speech. Hence, they contain complementary information, which is efficiently used for SLR, as demonstrated by studies on NIST LRE 2017 task.

The rest of the paper is organized as follows: Section 2 presents the significance of analytic phase in SLR using listening experiments. In Section 3, the feature extraction from long-time analytic phase for SLR is explained. The SLR experiments with NIST LRE 2017 database using different features are detailed in Section 4. Section 5, concludes the contributions of this paper to SLR.

*The authors would like to acknowledge the NUS Start-up grant for the project- 'Non-parametric approaches to voice morphing'.

2. SIGNIFICANCE OF ANALYTIC PHASE

The magnitude of speech is widely studied and utilized in digital processing of speech signals. However, there has not been equal attention given to phase of speech, possibly due to the phase wrapping problem, the nonlinear nature of solutions involving phase, and associated computational complexity [24]. It has been a challenge to derive stable and robust phase features from speech signals. But, phase play a significant role in conveying speaker and language specific information in speech signals.

The long-time analytic phase of a narrowband (NB) signal can be computed by constructing a complex analytic signal, employing Hilbert transform of the original signal and, finding the angle of the complex values [25]. If $s[n]$ is a discrete-time NB signal, the corresponding discrete-time equivalent of its analytic signal can be obtained as $z[n] = s[n] + js_H[n]$, where $s_H[n]$ is the Hilbert transformed signal, whose N-point discrete Fourier transform (DFT) is

$$S_H[k] = \begin{cases} 0, & k = 0, N/2 \\ S[k], & 1 \leq k \leq \frac{N}{2} - 1 \\ -S[k], & \frac{N}{2} + 1 \leq k \leq N - 1 \end{cases} \quad (1)$$

where $S[k]$ is the N-point DFT of $s[n]$ [26]. The $S[k]$ and $S_H[k]$ are combined to form the one-sided DFT sequence $Z[k]$, whose inverse DFT renders the complex discrete-time 'analytic' signal $z[n]$ [26]. The temporal amplitude and analytic phase of $z[n]$ are computed in terms of its real and imaginary parts, the $s[n]$ and $s_H[n]$ respectively, as

$$a[n] = \sqrt{s^2[n] + s_H^2[n]}, \quad \theta[n] = tan^{-1}\left(\frac{s_H[n]}{s[n]}\right) \quad (2)$$

The $a[n]$ and $\theta[n]$ can be deduced as amplitude and frequency modulations (AM-FM) existing in the NB signal $s[n]$ [25].

The NB demodulation and resultant AM and FM components cannot be interpreted directly for wideband (WB) speech signals. Hence a NB Gabor filter-bank is utilized to decompose the WB speech into multiple NB components and, discrete-time 'analytic' signals are computed. Thus long-time analytic phase for multiple NB components of speech can be obtained for feature extraction.

The multiband demodulation analysis (MDA) of WB speech signals [27] is illustrated in Fig. 1. A segment of speech shown in Fig. 1(a) is processed using NB Gabor filterbank (eg: Fig. 1(b)) to obtain multiple NB components. One such component filtered with a NB filter having center frequency of 400 Hz is shown in Fig. 1(c). The complex analytic signal is constructed for this NB component and, temporal amplitude & analytic phase are computed using Equation (2), and are shown in Fig. 1(d) and (e), respectively, thereby achieving multiband demodulation of WB speech signal.

The analytic phase plays a decisive role in conveying several valuable information embedded in speech signals, language-specific information being a prominent one among them. To demonstrate the effects of analytic phase on human perception of languages, we conduct listening tests to identify languages from speech samples with tampered analytic phase. The phase tampering is performed by replacing analytic phase of each NB component of speech with uniformly distributed random values $U[-\pi, \pi)$, as described in [24].

Speech samples from the training dataset of NIST LRE 2017 database are used for the listening test. The database consists of recordings grouped into five language clusters, and each cluster holds multiple dialects/variants of the language. We used the Chinese cluster that has Mandarin and Min, and the English cluster for British and General American English [28]. Five native speakers for each cluster were asked to listen to speech signals and identify the

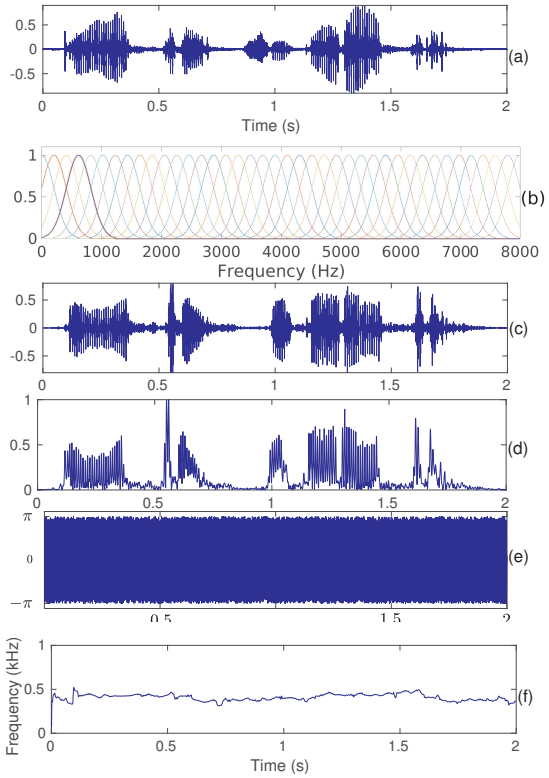

Fig. 1. Multiband demodulation analysis: (a) Speech signal, (b) Gabor filter-bank, (c) NB component (d) Temporal amplitude, (e) Analytic phase and (f) smoothed IF.

Table 1. Human language identification accuracy (%).

Type of speech	Chinese	English
Analytic phase-tampered	52	64
Original	94	96

language spoken as one among the dialects/variants in the respective cluster, or as 'can't say'. Each listener had listened to two sets of samples belonging to clean speech and analytic phase-tampered speech, each set having 10 samples with mean duration of 5 sec.

Table 1 reports the identification accuracy in the listening tests. It is evident from Table 1 that the absence of true analytic phase had significantly affected the ability of native speakers of languages in distinguishing their mother tongue from other similar sounding languages. Mandarin and Min are very different variants of Chinese language, where as British and General American are two accents of English language. Therefore, the identification accuracy in Chinese cluster is lower than that in English cluster. Also, the listeners had noted that they could differentiate between dissimilar languages (Chinese from English and vice versa) from analytic phase-tampered speech samples. Thus, analytic phase plays a considerable role in differentiating similar sounding languages, though not crucial in distinguishing very dissimilar ones, in human perception of languages from speech signals. It can make positive contributions in automatic identification of dialects/accents of languages.

As the perceptual significance of analytic phase in conveying language-specific information in speech signals is successfully demonstrated by the listening tests, we proceed to extract features

from long-time analytic phase for SLR.

3. FEATURE EXTRACTION FOR SLR

The computation of analytic phase of NB components of speech for feature extraction is not an unambiguous procedure. The phase obtained using inverse tangent as given in Equation (2), is the wrapped phase. All phase values are constrained to the interval $[-\pi, \pi)$ and any value of analytic phase falling outside this interval will be wrapped back into the interval itself (See Fig. 1(e)). Hence, unambiguous computation of analytic phase for any specified instant of time is not possible.

We use the IF, which is the time-derivative of the long-time analytic phase, for feature extraction. The computation of IF can be realized without explicitly computing and differentiating the analytic phase values. The values for IF at each instant can be calculated by differentiating the logarithm of analytic signal and equating the imaginary parts [24]. Also, the differentiation of complex analytic signal can be realized using the 'differentiation property' of Fourier transform. Thus IF of a discrete-time NB signal can be computed without getting affected by phase wrapping problem as [24],

$$\theta'[n] = \frac{2\pi}{N}\mathrm{Re}\left\{\frac{\mathscr{F}^{-1}(kZ[k])}{\mathscr{F}^{-1}(Z[k])}\right\}, \quad (3)$$

where $\mathscr{F}^{-1}$ denotes inverse DFT operation and $Z[k]$ is the N-point DFT of the analytic signal $z[n]$ corresponding to the NB signal $s[n]$. The operator $\mathrm{Re}\{.\}$ denotes the real part of a complex value.

The IF thus computed for the NB component of speech shown in Fig. 1(c) is smoothed to remove spurious variations using a moving average filter of 25 ms duration and, is given in Fig. 1(f). As the NB component under consideration is obtained by filtering the speech signal through a NB filter with center frequency of 400 Hz, the IF is centered around 0.4 kHz in the Fig. 1(f). Thus the IF can be interpreted as the frequency of a sinusoid which locally fits the NB signal and, exhibits the instantaneous variations in the frequency of the NB signal around the center frequency [24].

In order to illustrate the information captured by the IF of NB components of speech, we plot the scatter plot of all IF from all bands in the WB speech signal and is termed as the pyknogram [27]. The spectrogram of a segment of speech signal and corresponding pyknogram are shown in Fig. 2. The formant tracks in the speech spectrogram is replicated with high clarity in the pyknogram, revealing the efficacy of IF from NB components of speech in capturing linguistic and speaker-specific characteristics in speech signals. Also, the IFs capture the long-time instantaneous variations in speech characteristics and hence, can be beneficial to SLR.

Now we can perform feature extraction from long-time IF contours. The parameters of the filter-bank used to decompose WB speech into NB components have significant effects on IF computation and subsequent feature extraction. In this paper, we use Gabor filter-bank, as the roll-off rate of filter responses is smooth and hence its contributions to discontinuities in filter outputs are negligible [24, 27]. The number of filter channels and bandwidth of individual filters are other parameters requiring further investigation. The choice of filter bandwidth is a trade-off between the capturing of finer variations in IF around center frequency (not becoming extremely NB), without forcing the filter output to become WB. In this study, we used Gabor filters with a bandwidth of 400 Hz satisfying the considerations discussed [24, 27]. The number of channels in the filter-bank should cover the entire spectral range of the WB speech signal and, ensure that there exist an overlap of at least 50% between

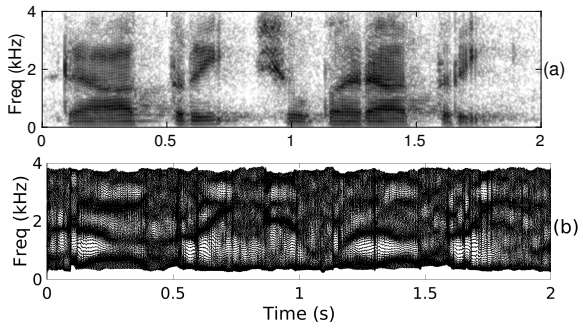

Fig. 2. (a) Spectrogram and (b) Pyknogram corresponding to the segment of speech shown in Fig. 1(a).

adjacent filter spectral responses to guarantee that the constant-Q criteria for exact reconstruction of a signal is satisfied [24]. In this study, we used 40 channels in the filter-bank, where each filter is centered at multiples of 100 Hz covering the entire spectral range (0-4 kHz) of WB speech signals sampled at 8 kHz sampling frequency. 40 IF contours were extracted from the NB components of speech and feature extraction is performed.

The long-time IF contours are segmented into short-time overlapping frames and are averaged within each frame, thereby obtaining 40 IF coefficients. As there exist 75% overlap between adjacent filter responses in the Gabor filter-bank, 50% overlap between second-neighboring filter and 25% overlap between third-neighbors, there exist a large amount of redundancy among the IF coefficients. Hence we apply discrete cosine transform (DCT) to decorrelate the IF coefficients and saves the first few coefficients, termed as IF cepstral coefficients (IFCC). The IFCC, together with delta and acceleration coefficients form the features for each short-time frame, that we call IFCC features.

The features from IF of NB components of speech were previously studied in speaker and speech recognitions [24, 29–31]. The IF computation reported in [30, 31] uses energy/amplitude weighting of IF to deal with spurious effects. But, this strategy affects the true nature of IF by masking and dominating the IF characteristics at higher temporal amplitudes. The IF computation reported in [29] relies on direct differentiation of analytic phase, which may be affected by phase wrapping problem. The IFCC features reported in this paper follow Hilbert transform demodulation and phase wrapping problem of analytic phase does not adversely affect the IF computation. Features from short-time phase spectrum (modified group delay) were also utilized for SLR [23]. As these features do not capture long-time information in speech signals, they may not be considerably beneficial to SLR. In this paper, we study the IFCC features, capturing long-time information in speech for SLR.

3.1. SDCC and IFCC

We propose to use the SDCC and IFCC features for SLR, capturing long-time information from spectral magnitude and analytic phase of speech signal, respectively. The SDCC are computed by windowing in time-domain to learn short-time spectral characteristics, whereas IFCC are computed by windowing in spectral domain to learn long-time temporal characteristics of speech signals. For SDCC extraction, the speech signals are segmented into short-time frames and the short-time Fourier transform (STFT) is applied. The MFCC features are extracted from STFT spectrum and the delta values between MFCC features from multiple neighboring frames are concatenated

to form SDCC features. Thus the SDCC are computed by short-time processing and feature extraction from speech, followed by capturing inter-frame relationships between the short-time features to obtain long-time information. On the contrary, IFCC features are obtained by long-time NB AM-FM demodulation of speech and associated computation of IF contours. Later, the long-time IF contours are segmented to produced short-time features termed as the IFCC. Thus, the SDCC and IFCC contain long-time information from two different domains. We have good reason to believe that they contain complementary language-specific information. We perform SLR trials with these features to study the information contained in them.

4. EXPERIMENTAL EVALUATION

4.1. Description of database

The training dataset released by NIST LRE 2017 was utilized for building the SLR system, following the rules for the *fixed* condition training. The dataset consists of speech recordings of 14 languages from 5 language clusters namely, Arabic, Chinese, English, Spanish, and Slavic, chosen from previous LRE data, Fisher corpus and Switchboard corpora [28]. We have chosen LRE 2017 development data (DEV17) and LRE 2015 evaluation data (EVAL15) as test datasets. The speech samples in test set are of conversational telephone speech - broadcast NB speech (MLS14) and video speech (VS). The MLS14 dataset consists of speech samples of duration 3, 10 and 30 seconds, whereas VS dataset has duration of entire video recording [28]. A total of 3,661 and 164,334 SLR trails are there in DEV17 and EVAL15 datasets, respectively.

4.2. Description of features and SLR system

All speech samples in this study are sampled at 8 kHz. The speech signals are segmented into frames of 25 ms duration, shifted by 10 ms. 9-dimensional MFCC are extracted from each short-time frame. The SDCC are extracted by computing delta coefficients between MFCC features over 1 neighboring frame before and after the current frame, and stacking these delta coefficients over 7 adjacent blocks, forming 63-dimensional SDCC features.

The filter-bank analysis of speech signals is carried out using a 40-channel filter-bank having filter bandwidth of 400 Hz and long-time IF contours are computed. These IF contours are segmented into frames of 25 ms duration, shifted by 10 ms and 60-dimensional IFCC features (20 IFCC+Δ + $\Delta\Delta$) are extracted. Energy based voiced activity decisions are computed for feature selection, and cepstral mean normalization of features is performed.

The SLR system in this study is a universal background model (UBM) - i-vector system [10,11], consisting of a 2048-mixture UBM and 600-dimensional total variability matrix. The channel compensation and scoring were done using a Gaussian back-end with linear discriminant analysis (LDA). The scores calibration and fusion were performed using FoCal toolkit by training a multiclass linear logistic regression (MCLR) model [32]. All the models were trained using the training dataset described in Section 4.1. The performance metric for evaluation of the SLR system is the average cost function based on two sets of values for costs of miss detections and false alarm, and apriori probability for target languages. The minimum value of average cost function ($minC_{avg}$) computed using detection thresholds that minimizes detection cost function will be reported as performance metric in this paper [28]. Also, the equal error rate (EER) will be reported for performance comparisons.

Table 2. Evaluation of acoustic features for SLR in terms of percentage of EER and $minC_{avg}$ (reported within parenthesis).

Features	DEV17		EVAL15
	MLS14	VS	MLS14
SDCC	10.22 (0.359)	6.49 (0.216)	11.82 (0.421)
IFCC	11.41 (0.374)	12.58 (0.421)	15.51 (0.501)
DBN	5.97 (0.218)	4.08 (0.143)	6.75 (0.249)
SDCC+IFCC	**7.15 (0.251)**	**5.32 (0.188)**	**9.44 (0.340)**
DBN+IFCC	**4.60 (0.166)**	**3.42 (0.129)**	**5.97 (0.222)**

4.3. Evaluation

The performance of SLR systems with SDCC and IFCC features were evaluated with two datasets, namely DEV17 and EVAL15. The EER and $minC_{avg}$ computed after evaluating the SLR systems over the entire set of trials in test datasets are reported in Table 2. The SDCC features perform better than the IFCC features in SLR. We noted that SDCC features are extracted from information in short-time spectral magnitude, while IFCC features are from long-time analytic phase, and expect them to contain complementary information. To explore this possibility, we attempt fusion of scores from subsystems based on SDCC and IFCC features (SDCC+IFCC) at the back-end using FoCal multiclass toolkit [32]. The fusion of information from SDCC and IFCC features had outperformed the SLR performance from SDCC alone in terms of both EER and $minC_{avg}$. In fact, the fusion delivered a relative improvement of 23.23% in average EER over all datasets in comparison with SDCC features.

We explore the possibility of complementary information in IFCC features to another set of features representing long-time information from spectral magnitude, namely the DBN features [15–18]. They are obtained from the bottleneck layer having 64 hidden neurons with ReLU activation function, in a DNN with 7 hidden layers of configuration $2520-1024\_5-64-1024-6111$. The inputs to the DNN are a stack of MFCC features from 10 to 20 frames of speech and, output nodes are set to predict senones from context-dependent HMMs. We note that DBN features also come from short-time spectral magnitude and hence, IFCC features may offer complementary information to them. The DBN+IFCC fusion delivered a relative improvement of 16.78% in average EER over all datasets, upon the DBN features alone. Hence the utilization of information from IFCC features combined with features from magnitude spectrum of speech signals (SDCC and DBN) indicated the existence of complementary language-specific cues in them, yielding better SLR performance.

5. CONCLUSIONS

We have explored the prominence of long-time analytic phase in representing language-specific information in speech signals. The listening tests conducted in our study revealed the significant effects of analytic phase on human perception of closely sounding languages. To exploit the valuable information contained in long-time analytic phase for SLR, we propose to extract the IFCC features from long-time IF contours, representing analytic phase. The IFCC features, in combination with SDCC and DBN features, were utilized for SLR using NIST LRE datasets. The experiments on SLR exhibited the existence of complementary information in features extracted from long-time information in spectral magnitude and analytic phase of speech signals. Thus, the language-specific information in long-time analytic phase can be effectively utilized for the betterment of state-of-the-art SLR systems using baseline features.

6. REFERENCES

[1] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013.

[2] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. on Spch and Audio Proc.*, vol. 4, no. 1, pp. 31–, Jan 1996.

[3] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, "Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation," in *Proc. Odyssey'06*, June 2006, pp. 1–8.

[4] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," in *Proc. ICSLP*, 2004, pp. 1283–1286.

[5] R. Tong, B. Ma, H. Li, and E. S. Chng, "A target-oriented phonotactic front-end for spoken language recognition," *IEEE Trans. on Audio, Spch., and Lang. Proc.*, vol. 17, no. 7, pp. 1335–1347, Sept 2009.

[6] L. F. D'Haro, R. Cordoba, C. Salamea, and J. D. Echeverry, "Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition," in *Proc. ICASSP*, May 2014, pp. 5342–5346.

[7] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "On the projection of PLLRs for unbounded feature distributions in spoken language recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1073–1077, Sept 2014.

[8] F. Allen, E. Ambikairajah, and J. Epps, "Language identification using warping and the shifted delta cepstrum," in *2005 IEEE 7th Workshop on Multimedia Signal Processing*, Oct 2005, pp. 1–4.

[9] W. Q. Zhang, L. He, Y. Deng, J. Liu, and M. T. Johnson, "Time-frequency cepstral features and heteroscedastic linear discriminant analysis for language recognition," *IEEE Trans. on Audio, Spch., and Lang. Proc.*, vol. 19, no. 2, pp. 266–276, Feb 2011.

[10] D. Martnez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in i-vectors space," in *Proc. Interspeech*, 2011.

[11] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proc. Interspeech*, 2011.

[12] H. Wang, C. C. Leung, T. Lee, B. Ma, and H. Li, "Shifted-delta MLP features for spoken language recognition," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 15–18, Jan 2013.

[13] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2, pp. 210 – 229, 2006.

[14] A. R. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. ICASSP*, May 2011, pp. 5060–5063.

[15] P. Matejka, L. Zhang, T. Ng, S. H. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," in *Proc. Odyssey*, 2014.

[16] Y. Song, R. Cui, X. Hong, I. Mcloughlin, J. Shi, and L. Dai, "Improved language identification using deep bottleneck network," in *Proc. ICASSP*, April 2015, pp. 4200–4204.

[17] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, Oct 2015.

[18] K. A. Lee, H. Li, L. Deng, et al., "The 2015 NIST language recognition evaluation: The shared view of I2R, Fantastic4 and SingaMS," in *Interspeech*, 2016, pp. 3211–3215.

[19] B. Yin, E. Ambikairajah, and F. Chen, "Combining cepstral and prosodic features in language identification," in *Proc. ICPR*, 2006, vol. 4, pp. 254–257.

[20] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, no. 10, pp. 782 – 796, 2008.

[21] D. Martnez, E. Lleida, A. Ortega, and A. Miguel, "Prosodic features and formant modeling for an i-vector-based language recognition system," in *Proc. ICASSP*, May 2013, pp. 6847–6851.

[22] R. W. M. Ng, T. Lee, C. C. Leung, B. Ma, and H. Li, "Spoken language recognition with prosodic features," *IEEE Trans. on Audio, Spch., and Lang. Proc.*, vol. 21, no. 9, pp. 1841–1853, Sept 2013.

[23] F. Allen, E. Ambikairajah, and J. Epps, "Warped magnitude and phase-based features for language identification," in *Proc. ICASSP*, May 2006, vol. 1, pp. I–I.

[24] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, no. Supplement C, pp. 54 – 71, 2016.

[25] L. Cohen, *Time-frequency analysis: theory and applications*, Signal processing series. Prentice Hall, Inc., Upper Saddle River, NJ, USA., 1995.

[26] S.L. Marple Jr., "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Sig. Proc.*, vol. 47, no. 9, pp. 2600–2603, Sep 1999.

[27] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The JASA*, vol. 99, no. 6, pp. 3795–3806, 1996.

[28] NIST, *NIST 2017 Language Recognition Evaluation Plan*, May 2017.

[29] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Speaker identification using FM features," in *11th Australian International Conference on Speech Science and Technology*, 2006, pp. 148–152.

[30] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. on Audio, Spch. and Lang. Proc.*, vol. 16, no. 6, pp. 1097–1111, Aug 2008.

[31] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency," *Speech Communication*, vol. 53, no. 5, pp. 707–715, May-Jun 2011.

[32] N. Brümmer, "FoCal Multi-class Toolkit," *http://niko.brummer.googlepages.com/focalmulticlass*, 2014.