INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Native Language Identification Using Spectral and Source-Based Features

Avni Rajpal[1]*, Tanvina B. Patel*[1]*, Hardik B. Sailor*[1]*, Maulik C. Madhavi*[1]*, Hemant A. Patil*[1] *and Hiroya Fujisaki*[2]

[1] Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), India
[2]University of Tokyo, Japan

{avni_rajpal, tanvina_bhupendrabhai_patel, sailor_hardik, maulik_madhavi, hemant_patil}@daiict.ac.in fujisaki@alum.mit.edu

Abstract

The task of native language (L1) identification from non-native language (L2) can be thought of as the task of identifying the common traits that each group of L1 speakers maintains while speaking L2 irrespective of the dialect or region. Under the assumption that speakers are L1 proficient, non-native cues in terms of segmental and prosodic aspects are investigated in our work. In this paper, we propose the use of longer duration cepstral features, namely, Mel frequency cepstral coefficients (MFCC) and auditory filterbank features learnt from the database using Convolutional Restricted Boltzmann Machine (ConvRBM) along with their delta and shifted delta features. MFCC and ConvRBM gave accuracy of *38.2%* and *36.8%*, respectively, on the development set provided for the ComParE 2016 Nativeness Task using Gaussian Mixture Model (GMM) classifier. To add complementary information about the prosodic and excitation source features, phrase information and its dynamics extracted from the log(F_0) contour of the speech was explored. The accuracy obtained using score-level fusion between system features (MFCC and ConvRBM) and phrase features were *39.6%* and *38.3%*, respectively, indicating that phrase information and MFCC capture complementary information than ConvRBM alone. Furthermore, score-level fusion of MFCC, ConvRBM and phrase improves the accuracy to *40.2*%.

Index Terms: Shifted delta cepstrum, Convolutional Restricted Boltzmann Machine, F_0, Accent, Phrase.

1. Introduction

In general, multilingual speakers lack thorough acquisition of second language (L2) and speech from a particular group of non-native speakers show common traits such as distinct 'foreign accent' and typical pronunciation errors [1], [2]. The task of native language (L1) identification aims at identifying such commonalities from spontaneous speech that can be used to identify the mother tongue of English (L2) speakers. The feasibility of proposed Native Language Identification (NLID) task in this challenge lies in phenomenon of *prosodic transfer* from L1 to L2. Moreover, the dialectal differences of learner's L1 in aspects of prosodic transfer from L1 to L2 should be known as observed by Fujisaki and others [3], [4]. NLID system can be used in parallel with computer-aided language learning systems to provide L1- specific training program, also for automated speech assessment systems, reading tutors, adaptation in ASR, speaker forensics etc. [5], [6].

Non-native speakers frequently maintain a foreign accent and inadvertently carry phonemic details from L1 to L2 [7]. In addition, non-native speech typically includes more disfluencies than native speech and is characterized by a lower speech rate [2], [8]. This indicates the influence of L1 over L2, both in terms of prosody as well as segmental aspects [1]. However, the degree of influence, depends on the amount of L1 used and the proficiency of L2 [2], [9]. The above difficulties are very prominent in Native Language Speech Corpus (NLSC), since the recordings are from the TOEFL iBT® assessment given by non-native speakers under examination conditions [10]. Thus, nativeness of speaker can be identified by studying the acoustic and prosodic aspects that remain native-like or become prominent while speaking L2.

In this paper, we intend to explore both acoustic and prosodic features for L1 identification. We propose to use acoustic features obtained from the auditory filterbank learnt from the speech signals using Convolutional Restricted Boltzmann Machine (ConvRBM). In majority of the experiments related to evaluation of degree of nativeness of the speaker and L2 (particularly English) acquisition, native English speaking listeners are used for human scoring [9], [11]. This is because, it is assumed that non-native cues can be easily identified by the native English speaking listeners that are having well-established linguistic knowledge of speech. This idea to imitate the hearing mechanism of native English speakers motivates us to learn features not only from NLSC but also from other databases, which contains speech recordings from native English speakers. In this work, we have used WSJ0 database having recording from clean environment and AURORA 4 multi condition database [12], [13].

The time interval or frame size for feature extraction is known to capture different segmental information as per the window chosen for analysis. For longer analysis window, the cepstral features contain information about formants structure and its dynamics that can reflect the movement and position of vocal and nasal articulators [14]. Moreover, a time interval of 90 ms is suggested by Furui [15] to preserve the transitional information associated with changes from one phoneme to another. Moreover, Soong and Rosenberg [16] proposed 100 to 160 ms time interval, to obtain good estimates of the trend of spectral transitions between the syllables. Thus, in order to extract segmental information, the proposed features for L1 identification task are extracted over longer window durations. Furthermore, the dynamic features (i.e., delta (Δ), delta-delta (ΔΔ)) and Shifted Delta Cepstral (SDC) features were explored to capture the spectral dynamics. To gain the advantage of prosodic cues for L1 identification task, prosodic features such as fundamental frequency (F0), phrase and accent along with their dynamics are considered. The results indicate that features from WSJ0 and AURORA 4 can better identify L1 than features from NLSC. Moreover, phrase with its dynamics was found to capture complementary information with respect to spectral features and was used to further improve the accuracy of the classification system.

http://dx.doi.org/10.21437/Interspeech.2016-1100

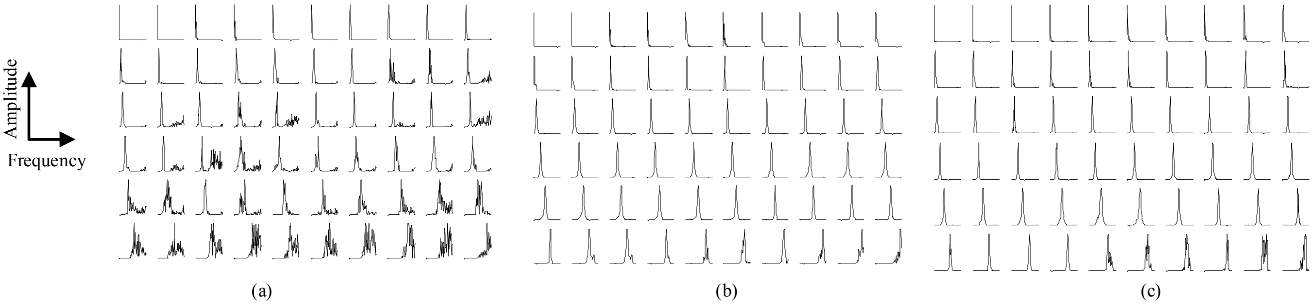

Figure 1: *Subband filters in frequency-domain trained on (a) NLSC, (b) WSJ0 and (c) AURORA 4.*

2. Acoustic Features

The NLID task requires features that capture idiosyncrasies of a particular L1 that remain while speaking L2. Thus, for this task, both spectral and excitation source features are explored.

2.1. Convolutional Restricted Boltzmann Machine

ConvRBM is an unsupervised probabilistic model used to learn auditory filterbanks directly from speech signals. The inference is based on sampling of hidden units from noisy rectified linear units (NReLU) [16]. Training is based on contrastive divergence and parameters are updated using gradient descent. The training of the model and feature extraction is similar to our very recent work reported in [16]. The model is trained using three databases, namely, NLSC, WSJ0 and AURORA 4. The subband filters learnt on these three databases are shown in Figure 1. To analyze the subband filters, they were arranged according to their center frequency. As shown in Figure 1, we can see that many subband filters trained on NLSC are different than the subband filters trained on native English speakers' databases (WSJ0 and AURORA 4). As seen from Figure 1(a) many of the filters are not localized specifically in mid-center frequency range compared to subband filters in Figure 1(b) and Figure 1(c). This indicates that speaker-specific traits are learnt by NLSC, which may be due to interference of L1 over L2 [18].

2.2. Shifted Delta Cepstral (SDC) Features

SDC are long-term temporal features that capture the spectral dynamics of speech via cepstral trajectory in *N*-dimensional (dim) feature space. SDC are known to improve the performance of speaker recognition and language recognition systems [14], [19], [20]. SDC has pseudo-prosodic behavior, since in each frame; it captures the temporal dynamics of the articulators present in the next frames [14]. For a given *N*-dim cepstral feature vector, SDC vector at n^{th} frame is obtained by concatenating *k* blocks of delta coefficients and is given by

$$\Delta c(n+iP) = \sum_{d=-D}^{D} dc(n+iP+d) \Big/ \sum_{d=-D}^{D} d^2 \quad , \qquad (1)$$

where *D* is time advance and delay for delta computation, *P* is time shift between consecutive blocks, *i*=0 to *k-1* blocks to be concatenated [14].

2.3. Excitation Source-Based Parameters

The F0 contour of speech is known to capture both linguistic and non-linguistic information (speech prosody). This information is embedded in the low frequency variations (LFV) and high frequency variations (HFV) of the F0 contour [21]. The state-of-the-art Fujisaki model decomposes F0 in log- domain into phrase and accent components extracted from LFV and HFV, respectively [21], [22]. In this work, instead of extracting components from their respective commands, we use the LFV and HFV directly along with their dynamics for NLID task. We have used the Zero Frequency Filtering (ZFF) method to estimate the F0 contour [23]. The negative-to-positive zero-crossings of the zero frequency filtered signal gives an estimate of the Glottal Closure Instant (GCIs) Thereafter, the F0 contour is obtained from the GCI locations. In order to estimate HFV and LFV, log (F0) is processed, i.e., the intermediate values of log (F0) for unvoiced speech regions and short pauses are interpolated and microprosodic variations due to individual speech sounds (such as plosive, fricatives, etc.) are smoothed out. The interpolated and spline fitted log (F0) contour is then passed through a highpass filter with a stop frequency at 0.5 Hz to obtain HFV. The HFV is then subtracted from the processed log (F0) contour yielding LFV. LFV consists of phrase component and Fb (speaker-specific constant) [21], [24]. Fb is set to the overall minimum of the LFV and is subtracted from it to give phrase component.

3. Experimental Results

3.1. Speech Corpus

NLSC is a corpus of non-native English speech consisting of spoken response provided during a high-stakes global assessment of English language proficiency, the Test of English as a Foreign Language (TOEFL iBT®). It contains *5,132* spoken responses. For the ComParE 2016 Nativeness Task, the NLSC corpus is partitioned as follows: *3,300* responses (*64 %,* approximately *41.3* hours) will be used as training data, *965* responses (*19 %,* approximately *12.1* hours) will be used as development data, and *867* responses (*17 %*, approximately *10.8* hours) will be used as the test data. The complete details of the corpus are mentioned in [10].

3.2. System Building

In this paper, we use Gaussian Mixture Model (GMM) with *128* mixtures for performing the classification of input speech into given L1 classes. At the training stage, *11* GMM models each corresponding to separate L1 class were built. The models were trained using *300* instances of each L1 class available from the training set. Final scores are represented in terms of log-likelihood (LLK). The decision for the predicted class for the test speech is taken by calculating LLK for each GMM class. The class associated with the GMM that has maximum LLK is the predicted class.

3.3. Performance Measure

In this paper, Unweighted Average Recall (UAR) and weighted average recall (WA) ('conventional accuracy') is used as the evaluation measure [10]. UAR is given by [25]:

$$UAR = \frac{1}{N}\sum_{i=1}^{N}(C_{ii} \Big/ \sum_{j=1}^{N} C_{ij}) \,, \quad (2)$$

where C_{ij} is the number of instances of class i in contingency matrix C that are classified as class j with N as the total number of the classes. To utilize complementary information, score-level fusion of features is obtained using (3).

$$LLK_{fused} = \sum_{i=1}^{N} \alpha_i \, LLK_{feati} \,, \quad (3)$$

where LLK_{feati} is the LLK score of i^{th} feature, α_i decides the weight of the scores such that $\sum_{i=1}^{N} \alpha_i = 1$. Before score-level fusion, the scores were normalized using *tanh-estimators* [26]. This technique was chosen for normalization since it is robust and is not sensitive to the outliers [26]. The normalization is given by:

$$LLK'_k = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{LLK_k - \mu}{\sigma}\right)\right) + 1\right\} \quad , \quad (4)$$

where LLK_k and LLK'_k are the original and normalized score of the k^{th} class, μ and σ are the global mean and standard deviation of the scores, respectively.

3.4. Effect of Window Length of Spectral Features

In this work, we consider *13*-dim NLSC, WSJ and AURORA features. In addition, we also used *13-dim* MFCC as spectral features. To capture the short-time dynamic (DYN) information, these static spectral features are concatenated with Δ and ΔΔ to form *39-dim* MFCC+DYN, NLSC+DYN, WSJ+DYN and AURORA+DYN. Furthermore, to capture long-term temporal spectral dynamics, the *13 dim* spectral features were also combined with SDC resulting in *39-dim,* MFCC+SDC, NLSC+SDC, WSJ+SDC and AURORA+SDC features. The parameters for SDC are set *to N=13, D=2, P=2, k=2* in (1). The *13-dim* spectral features are extracted over various window durations (*25 ms, 100 ms, 150 ms* and *200 ms*). This is to quantify the fact that in spectral features, long duration window is useful in capturing the prosodic trends. Therefore, the window that is suitable for NLID task is investigated. As shown in Figure 2, WA on the development set increases with increase in window length. Moreover, *150 ms* and *100 ms* are ideal window lengths for both DYN and SDC features.

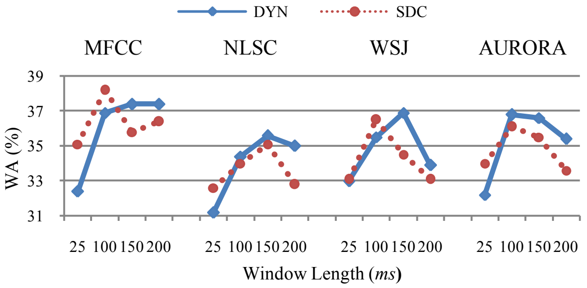

Figure 2: *Plot of window length (ms) vs. WA (in %).*

3.5. Effect of Dynamics of Source Features

In order to capture the temporal dynamic information contained in log(F_0), Phrase and HFV (denoted as Accent) features, their, Δ, ΔΔ and ΔΔΔ (3Δ) components were extracted. Figure 3 shows the effect of dynamics of source-based features on the classification accuracy. It can be observed from Figure 3 that dynamic features of log(F_0) and phrase contain significant information about L1 of the speakers. However, the dynamic information undermines the performance of accent features.

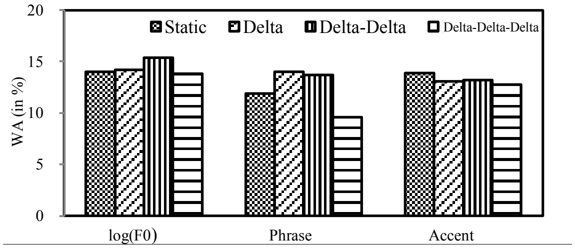

Figure 3: *Plot of dynamic features vs. classification accuracy for source features.*

3.6. Results on the Development Set

The best results from Figure 2 are summarized in Table 1 and it is observed that among data-driven approaches, WSJ and AURORA perform equally well with accuracy of *36.9 % and 36.8 %,* respectively, compared to the NLSC features. The possible reason of underperformance of NLSC could be that the speaker-specific traits learnt by the model were not specific to an individual group of non-native speakers, instead generalized statistical (invariant) properties corresponding to entire database was learnt. Hence, the models learnt on NLSC might not be optimal for individual non-native speaker group. However, the model trained on native English speakers' database represents an *optimal auditory code* [17], [27] that captured the common traits of non-native speakers. From Table 1, we also observe that the accuracy of handcrafted MFCC+SDC features is highest, i.e., it performs better than our proposed data-driven features (WSJ and AURORA) specifically with SDC. The possible reason could be that the model trained on WSJ0 and AURORA 4 does not incorporate any information from NLSC database. However, there is not huge difference in the performance of WSJ and AURORA, which could be improved in future by adapting models trained on WSJ0 and AURORA 4 to specific non-native speaker groups in NLSC database [28]. Among the source-based features, Table 1 shows that, log(F_0) has highest accuracy than phrase and accent features when used independently for classification. However, accuracy of log(F_0) feature significantly improves when accent component is fused with it at feature-level. Overall, it can be observed from Table 1 that the prosodic features based on Fujisaki model alone are not performing well for the proposed NLID task. Authors believe that this is primarily due to lack of dialectal knowledge about L1 speaker and its prominent effect on L2 [3], [4].

Table 1: *WA and UAR of spectral and source features*

Spectral* Features	WA (%)	UAR (%)	Source Features	WA. (%)	UAR (%)
MFCC+DYN_150	37.4	37.7	logF_0	14.0	14.1
MFCC+SDC_100	**38.2**	**38.4**	Phrase	11.9	11.9
NLSC+DYN_150	35.6	35.8	Accent	13.9	13.9
NLSC+SDC_150	35.1	35.2	logF_0+Phrase	14.1	14.1
WSJ+DYN_150	**36.9**	**37.1**	logF_0+Accent	**14.9**	**15.1**
WSJ+SDC_100	36.5	36.6	Phrase+Accent	13.1	13.0
AURORA+DYN_100	36.8	**37.1**	logF_0+Phrase+Accent	14.8	14.8
AURORA+SDC_100	36.1	36.3			

* MFCC+DYN_150: Feature_Window Length (*ms*).

The best performing data-driven features (WSJ+DYN_150 and AURORA+DYN_100) are fused with the spectral MFCC and source-based features. Among all the source-based features, i.e., log(F_0), phrase and accent (with all its combination and dynamic features), phrase+3Δ showed the improvement in accuracy of spectral features after score-level fusion. The selected features for which the better performance is obtained after fusion are shown in Table 2.

Table 2: *WA and UAR (in %) after score-level fusion*

Spectral Features	Fused Feature	α_2	WA (%) after fusion	UAR (%) after fusion
AURORA+DYN_100	MFCC+SDC_100	**0.6**	**39.6**	**39.8**
WSJ+DYN_150	MFCC+SDC_100	0.6	38.9	39
AURORA+DYN_100	Phrase+ 3 Δ	0.2	**38.3**	**38.5**
WSJ+DYN_150	Phrase+ 3 Δ	0.1	37.7	37.9

From Table 2, it can be observed that after fusion WA of both AURORA and WSJ features significantly improved. However, the improvement was more significant for AURORA even though they performed equally well without fusion. The possible reason is that AURORA 4 contains utterances from clean, noisy and mismatched conditions similar to NLSC database in which mismatched conditions was observed for few utterances. Thus, the model learnt using AURORA 4 is more generalized compared to the clean WSJ database. Thus, both at spectral-level and source-level, AURORA+DYN_100 is the common feature which give best fusion accuracy with MFCC+SDC_100 and phrase+3Δ. To further analyze the reason on improvement after fusion with MFCC+SDC_100 and phrase+3Δ features, the individual performance of the languages is considered (as shown in Table 3).

Table 3: *Effect of fusion on accuracy of individual L1*

	MFCC+ SDC_100	AURORA + DYN_100	AURORA +Phrase +3 Δ	AURORA+DYN_100+MFCC+ SDC_100	AURORA+DYN_100 + MFCC+SDC_100+ Phrase3 Δ
ARA	38	36	31	**41**	**37**
CHI	54	55	51	**55**	54
FRE	30	28	**45**	**28**	**38**
GER	**52**	48	46	**55**	**56**
HIN	33	34	37	**42**	**36**
ITA	**44**	37	31	**43**	**39**
JPN	34	38	36	35	35
KOR	41	44	38	42	43
SPA	**30**	19	**35**	**29**	**34**
TEL	**42**	40	36	**40**	**42**
TUR	**25**	29	**37**	28	**29**

It was observed from Table 3 that after fusion with MFCC, the accuracy of all languages showed improvement (except JPN, KOR and TUR). On the contrary, phrase+3Δ showed significant improvement for FRE, SPA and TUR languages. This indicates that phrase+3Δ carry significant information about nativeness of the FRE, SPA and TUR speakers. In order to exploit the advantage of these observations, all the three features (AURORA+DYN_100, MFCC+SDC_100 and Phrase+3Δ) were further fused using (3). The weighted combination of α_1*=0.5*, α_2*=0.4* and α_3*=0.1* gave relatively best WA *40.2 %* and UAR *40.4 %* for the challenge.

3.7. Result on Test set

The results of the test are shown for *2* out of the *5* trials that can be submitted for the challenge. The first trial include scores submitted for the feature MFCC+SDC_100 and for the second trial, scores from the fusion of AURORA, MFCC and phrase-based features (that performed best on development set) were submitted. The Confusion Matrix (CM) for first and second trial is shown in Table 4 (a)-4(b), respectively, sorted in geographical order from west to east. We obtain WA as *31.0 %* and *34.1 %,* UAR as *31.4 %* and *34.3 %* for first and second trial, respectively. The last row of CM indicates the accuracy on individual L1. The grey portions indicate higher confusion between languages. Table 4(a) shows that for MFCC+SDC_100, majority of languages were confused with ARA and CHI irrespective of geographical locations. However, the confusion for ARA is reduced in second trial; this is reflected with higher accuracy for all languages in second trial (except ARA and ITA). This may be due to the complementary information added due to AURORA+ DYN_100 and phrase+3Δ.

Table 4 (a): *The CM of test set for the first trial submission.*

Reference

		GER	FRE	ITA	SPA	ARA	TUR	HIN	TEL	JPN	KOR	CHI
Hypothesis	GER	**21**	2	4	4	2	4	4	1	1	3	3
	FRE	4	**12**	3	2	1	2	1	1	0	0	0
	ITA	11	9	**23**	4	6	2	1	3	6	3	3
	SPA	6	9	6	**23**	4	5	4	8	2	5	5
	ARA	8	17	4	11	**32**	24	10	15	8	8	10
	TUR	2	2	0	1	0	**9**	0	0	1	3	1
	HIN	0	2	1	4	2	1	**15**	11	1	1	4
	TEL	3	5	5	5	7	13	28	**39**	1	2	2
	JPN	0	2	2	4	8	3	2	2	**36**	14	6
	KOR	1	5	3	4	4	5	5	1	5	**23**	4
	CHI	19	13	17	15	14	22	12	7	14	18	**36**
	Acc(%)	28	15.4	33.8	29.8	40	10	18.3	31.8	48	28.8	48.6

Table 4 (b): *The CM of test set for the second trial submission.*

Reference

		GER	FRE	ITA	SPA	ARA	TUR	HIN	TEL	JPN	KOR	CHI
Hypothesis	GER	**26**	8	3	5	2	5	4	1	1	2	4
	FRE	6	**20**	6	5	1	4	3	0	1	2	1
	ITA	8	8	**20**	6	6	4	1	1	4	4	3
	SPA	6	9	10	**23**	9	12	4	6	5	5	6
	ARA	6	10	6	10	**29**	18	12	11	8	7	8
	TUR	3	1	0	2	2	**14**	0	1	0	4	1
	HIN	0	0	0	2	4	1	**17**	11	0	1	4
	TEL	1	5	3	2	4	9	22	**43**	0	0	3
	JPN	0	1	2	3	10	3	2	2	**42**	12	4
	KOR	1	5	2	5	3	4	6	6	4	**26**	4
	CHI	18	11	16	14	11	16	11	6	10	17	**36**
	Acc(%)	34.6	25.6	29.4	29.9	36.3	15.5	20.7	48.9	56	32.5	48.6

It was observed on the development set that, fusion of AURORA, MFCC and phrase-based features performed better for GER and FRE than MFCC only. This is reflected in test set as the confusion of GER and FRE to other languages reduced. Moreover, for both first and second trial majority of the languages are confused with CHI, which was not the case for development set. Thus, the fusion factor obtained over the development set did not generalize for the test set. In future, the effect on accuracy by taking the increased weightage of prosodic features for fusion will be explored.

4. Summary and Conclusions

Native speakers have linguistic knowledge of L1 that help in identifying non-native cues of L2 speakers. With respect to this idea, we attempt to learn features from native English speakers' database, i.e., WSJ0 and AURORA 4. Moreover, we also explored segmental information in the form dynamic and shifted delta features. The prosodic information from phrase and its dynamic features was found useful for NLID. On the test, among all languages CHI and JPN were found to perform better than the rest and the performance of TUR was low for all feature set. Our future research work will be directed towards finding prosodic and segmental cues motivated from the listener's perspective.

5. References

[1] J. Lopes, I. Trancoso, and A. Abad, "A nativeness classifier for TED talks," in Proc. Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic, pp. 5672–5675, 2011.

[2] C. Bergmann, S.A. Sprenger, and M.S. Schmid,"The impact of language co-activation on L1 and L2 speech fluency", Acta Psychologica,vol. 161, pp. 25-35, 2015.

[3] H. Fujisaki and M. Sugito, "Word accent and sentence intonation in foreign language learning," In Preprints of Papers for the Working Group on Intonation, Edited by Hiroya Fujisaki and Eva Garding, The *13*th Int'l Congress of Linguists, Tokyo, pp.109-119, 1982.

[4] L. Rasier and P. Hiligsmann, "Prosodic transfer from L1 to L2. Theoretical issues," Nouveaux Cahiers De Linguistique Française, vol. 28, pp.41-66, 2007.

[5] Min Ma, Keelan Evanini, Anastassia Loukina, Xinhao Wang, "Using F_0 Contours to Assess Nativeness in a Sentence Repeat Task", in Proc. INTERSPEECH, Dresden,Germany, pp. 653-657, 2015.

[6] S. Sam, X. Xiao, L. Besacier, E. Castelli, H. Li, and E. S. Chng, "Speech modulation features for robust non-native speech accent detection," in Proc. INTERSPEECH, Florence, Italy, pp. 2417-2420, 2011.

[7] P. Macizo, "Phonological coactivation in the bilinguals' two languages: evidence from the color naming task", Biling. Lang. Cogn., vol. 19, no. 2, pp. 361-375, 2016.

[8] J. van Doremalen, C. Cucchiarini, H. Strik, "Optimizing automatic speech recognition for low-proficient non-native speakers", EURASIP Journal on Audio, Speech, and Music Processing, vol. 2010, no.1, pp. 1-13, 2009.

[9] J.E. Flege, E.M. Frieda, T. Nozawa, "Amount of native-language (L1) use affects the pronunciation of an L2", Journal of Phonetics, vol. 25, no.2, pp. 169–186, 1997.

[10] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini,"The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language", in Proc. INTERSPEECH, ISCA, San Francisco, USA, 2016.

[11] Teixeira et al., "Evaluation of speaker's degree of nativeness using text-independent prosodic features," in Proc. of the Workshop on Multilingual Speech and Language Processing, Aalborg, Denmark, 2001.

[12] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in Proc. of the Workshop on Speech and Natural Language, Stroudsburg, PA, USA, HLT '91, pp. 357–362, Association for Computational Linguistics (ACL), 1992.

[13] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Tech. Rep., Inst. for Signal and Information Process, Mississippi State University, 2002.

[14] D.R. González and J. R. Calvo de Lara. "Speaker verification with shifted delta cepstral features: Its Pseudo-Prosodic Behaviour," in Proc. I Iberian SLTech 2009.

[15] S. Furui, "Cepstral analysis for automatic speaker verification", IEEE Trans on Acoustics, Speech, and Signal Processing, vol 29, no. 2, pp. 254-272, 1981.

[16] F. Soong and A. Rosenberg. "On the use of instantaneous and transitional spectral information in speaker recognition." IEEE Trans on Audio Speech and Signal Proc. vol 36, no. 6, pp. 871-879, 1988.

[17] H. B. Sailor and H. A. Patil, "Filterbank learning using convolutional restricted boltzmann machine for speech recognition," in Proc. Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, pp. 5895-5899, 2016.

[18] M. S. Lewicki, "Efficient coding of natural sounds," Nature Neuroscience, vol. 5, no. 4, pp. 356–363, 2002.

[19] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in Proc. Int'l Conference on Spoken Language Processing (ICSLP), Colorado, USA, pp. 89-92, 2002.

[20] M. V. Segbroeck, R. Travadi and S. S. Narayanan. " Rapid language identification," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 7, pp. 1118-1129, 2015.

[21] H. Fujisaki, R. Tomana, S. Narusawa, S. Ohno, C. Wang, "Physiological mechanisms for fundamental frequency control in Standard Chinese," in Proc. Int'l Conference on Spoken Language Processing (ICSLP), Beijing, China, pp. 9–12, 2000.

[22] H. Fujisaki and S. Nagashima, "A model for synthesis of pitch contours of connected speech," *Annual Report, Engg. Res. Inst., University of Tokyo*, vol. 28, pp. 53–60, 1969.

[23] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," IEEE Trans. on Speech and Audio Process., vol. 16, no. 8, pp. 1602-1613, 2008

[24] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," in Proc. Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, vol. 1, pp. 1281-1284, 2000.

[25] A. Rosenberg,"Classifying Skewed Data: Importance Weighting to Optimize Average Recall." in Proc. INTERSPEECH, Oregon, USA, pp. 2242-2245, 2012.

[26] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," Pattern Recognition, vol. 38, no. 12, pp.2270–2285, 2005.

[27] E. Smith and M. S. Lewicki, "Efficient coding of time-relative structure using spikes," Neural Comput., vol. 17, no. 1, pp.19–45, 2005.

[28] Kashiwagi et.al,"Divergence estimation based on deep neural networks and its use for language identification," in Proc. Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, pp. 5435-5439, 2016.