Interspeech 2018
2-6 September 2018, Hyderabad

Anomaly Detection Approach for Pronunciation Verification of Disordered Speech using Speech Attribute Features

Mostafa Shahin[1], Beena Ahmed[1], Jim X. Ji[1], Kirrie Ballard[2]

[1]Dept.of Electrical and Computer Engineering, Texas A&M University, Doha, Qatar
[2]Faculty of Health Sciences, The University of Sydney, Sydney, Australia

Abstract

The automatic assessment of speech is a powerful tool in computer aided speech therapy for disorders such as Childhood Apraxia of Speech (CAS). However, the lack of sufficient annotated disordered speech data seriously impedes the accurate detection of pronunciation errors. To handle this deficiency, in this paper, we used the novel approach of tackling pronunciation verification as an anomaly detection problem. We achieved this by modeling only the correct pronunciation of each individual phoneme with a one-class Support Vector Machine (SVM) trained using a set of speech attributes features, namely the manner and place of articulation. These features are extracted from a bank of pre-trained Deep Neural Network (DNN) speech attributes classifiers. The one-class SVM model classifies each phoneme production as normal (correct) or an anomaly (incorrect). We evaluated the system using both native speech with artificial errors and disordered speech collected from children with apraxia of speech and compared it with the DNN Goodness of Pronunciation (GOP) algorithm. The results show that our approach reduces the false-rejection rates by around 35% when applied to disordered speech.

Index Terms: pronunciation verification, disordered speech, one class SVM, deep learning, speech attributes.

1. Introduction

Childhood Apraxia of Speech (CAS) is a neurological speech disorder that affects a child's ability to make accurate movements for speech. Children with CAS suffer from different types of pronunciation difficulties such as inappropriate prosody, articulatory struggling and inconsistent speech sound production [1]. In this work we focus on using phoneme-level pronunciation verification to automatically detect the inconsistencies in the child's production for use in remote speech therapy systems. However, the accuracy of the automatic methods used to verify the correctness of pronunciation is crucial, as inaccurate verification can lead to misleading feedback causing delay in the treatment.

A number of different approaches have been proposed to achieve pronunciation verification. The most widely used of these is lattice-based mispronunciation detection, which works by constructing a lattice with the correct pronunciation and the most common pronunciation errors. In [2], a specific phoneme-level lattice for each prompt word was generated using the correct phoneme sequence with expected mispronunciations added as alternatives for use in learning the pronunciation of Quranic Arabic. Similar approaches have been applied to speech therapy [3] and second language acquisition [4]. In our previous work [5] we enhanced the accuracy of a lattice-based pronunciation verification method for disordered speech by using a Deep Neural Network Hidden Markov Model (DNN-HMM) acoustic model instead of the traditional Gaussian Mixture Hidden Markov Model (GMM-HMM) acoustic model. Most recently, Li et al. [6] introduced the acoustic-graphemic-phonemic model (AGPM) by combining the acoustic features along with the graphemes and canonical transcription in one multi-distribution DNN model. However, these methods are effective only as long as the errors fall within the probable pronunciation variants in the search lattice; their performance is degraded when unexpected pronunciation errors occur. Another issue is that a large amount of mispronunciation data is needed to accurately model possible errors, which is usually infeasible.

Another approach measures the confidence score of the pronunciation and compares it to a threshold to decide whether the pronunciation is correct or not. In [7], the authors compared the correlation of three different scores, namely the log-likelihood score, the segment duration score and the log-posterior probability score, with manual assessments; the best accuracy was achieved with the posterior probability score. The most common score used however is the so-called Goodness of Pronunciation (GOP) introduced by Witt and Young [8]. This score is computed by estimating the phoneme-level posterior probability via the output of a phoneme-loop recognizer based on the HMM acoustic model. The GOP has become a de-facto standard for measuring pronunciation quality and implemented in a vast number of applications including disordered speech [9-12].

Phoneme-level pronunciation error detection has also been achieved using binary classifiers where each phoneme is classified as either "correct" or "incorrect". Support Vector Machine (SVM) binary classifiers have been used widely [13-16] and shown to outperform the likelihood-based methods. Other classifiers such as Linear Discriminative Analysis (LDA) and decision tree have also achieved acceptable accuracy [17]. Despite the significant improvements obtained by these classifiers over the GOP algorithm, they are highly dependent on the availability of sufficient annotated mispronunciations to form negative samples used in the training of the binary classifier. Moreover, human labeling of non-native speech data in general, and disordered speech specifically, is more challenging than native speech data [18, 19], which adds an additional source of error in the data used to train the mispronunciation detector.

In this paper, we use the novel approach of casting the phoneme-level pronunciation verification problem as an anomaly detection problem. To handle the lack of sufficient mispronounced training data, we propose modelling each phoneme with a One-Class Support Vector Machine (OCSVM). Using a one-class model instead of typical multi-class models allows each phoneme model to be trained using only correctly pronounced data. The input features to the OCSVM represent the manner and place of articulation attributes of the phoneme derived from a set of speech attribute detectors based on binary DNN classifiers. The OCSVM learns the distribution of the phoneme attributes and

10.21437/Interspeech.2018-1319

then evaluates any input unseen speech segment by measuring its similarity with the phoneme model and deciding if it is similar (normal) or dissimilar (anomaly). To demonstrate the effectiveness of the algorithm we applied it on two speech corpora collected from typically developing (TD) children and children with CAS and compared it to the DNN GOP algorithm, the most common and well-proven pronunciation verification technique.

2. Method

2.1. System overview

Figure 1 presents a flowchart of our proposed system. First, a set of acoustic features are extracted from the speech signal and then passed to the forced alignment module along with the expected phoneme sequence. We use DNN-HMM acoustic models trained with Mel-Frequency Cepstral Coefficients (MFCC) features plus the delta and acceleration coefficients to perform forced alignment. Forced alignment is used to determine the time boundaries of each phoneme. Each phoneme is then mapped to its corresponding speech attributes and used to train a bank of speech attribute DNN binary classifiers. The trained speech attributes classifiers are used to extract a speech attribute feature vector from each frame that represents the likelihood it belongs to each attribute.

The frames of each phoneme are then converted to their corresponding speech attribute features and then used to train a phoneme-specific OCSVM model. In the testing mode, each frame is evaluated as correctly pronounced if it is classified by the OCSVM as correct (in-class) and evaluated as incorrectly pronounced if detected as an anomaly (out-of-class)

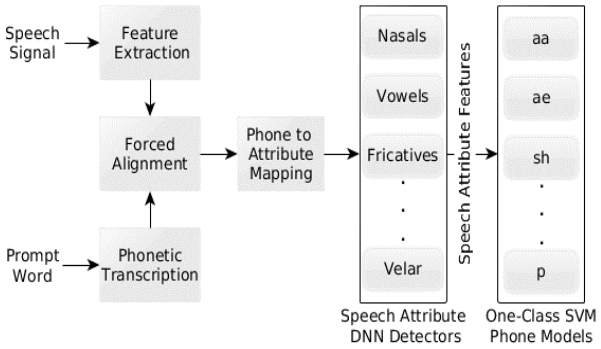

Figure 1: *The system flowchart.*

2.2. Speech attributes classifier

Each phoneme can be characterized by a set of attributes describing its manners and places of articulation. The detection of a phoneme's speech attributes is a well-known problem with different applications such as in, bottom-up automatic speech recognition system [20], identification of spoken language [21], lattice rescoring for LVSR [22] and universal phoneme recognition [23]. Recent work on speech attribute detection using DNNs has shown to significantly improve their performance [24].

In this paper we adopted 26 speech attributes as listed in Table 1 [20]. The phoneme-attribute mapping is adopted from [25]. A binary DNN classifier is trained to determine the existence or the absence of each individual attribute. The DNN consists of an input layer, output layer and a tunable number of hidden layers. The size of the input layer depends on the size of the input feature vector. From each frame of 25 msec length, we extracted a 78-feature vector consisting of 26 log filter bank energies along with their respective delta and acceleration coefficients. We then concatenated each of these vectors with four vectors on either side to form one super-vector input of size 702 that captures context variation. Finally, a two-way softmax layer lies on the top of the DNN estimating the posterior probabilities of the presence (+ve) and absence (-ve) of the attribute. Data from all phonemes sharing the same attribute are used as (+ve) samples while data from all other phonemes as (-ve) samples. We performed training using the mini-batch stochastic gradient descent method. A separate validation set was used to control the learning rate and the final accuracy reported using a different test set. To prevent biasing of the model we selected an equal number of (+ve) and (-ve) samples from the training, validation and testing datasets. The number of layers was tuned from 1 to 6 with a fixed number of nodes per layer (typically 2048).

Table 1: *List of speech attributes*

Vowels, Stops, Affricates, Fricatives, Nasals, Liquids, Semivowels, Approximant, Coronal, High, Dental, Glottal, Labial, Low, Mid, Velar, Back, Retroflex, Anterior, Continuant, Round, Tense, Voiced, Monophthongs, Diphthongs, Silence

2.3. One-class SVM

The OCSVM is a special variant of the traditional two class SVM introduced for the first time by Schölkopf et al. [26] to address the novelty detection problem. Unlike the multi-class SVM, here data from only one class is available and the OCSVM trained to create a decision boundary separating the data from the origin. The OCSVM is used commonly for anomaly detection applications [27]; in speech analysis, it has been used successfully to classify between speech and music [28], audio-event detection [29] and spoofing detection [30].

The OCSVM operates better when there are no or less anomalies in the training data, as in the pronunciation verification problem where more correctly pronounced than mispronounced data is available. This is because the decision boundary of the OCSVM is affected significantly by the existence of outliers [31]. Moreover, there are unlimited variations in the incorrect pronunciations, influenced by the speaker's native language in second language acquisition applications or the type and/or degree of disorder in speech therapy applications. Therefore, systems trained using the available mispronounced data fail to generalize to unseen variations of the pronunciation errors.

We trained an OCSVM model for each phoneme using only the correct pronunciation occurrences of this phoneme. The features used to train the OCSVM were extracted from the 26 pre-trained speech attributes classifiers described in Section 2.2, all of which have been found to be robust to speaker variation and environmental noise [8]. In addition, any distortion in the pronunciation of a specific phoneme is represented as a loss of one or more of its articulation attributes.

All the speech attribute binary classifiers were used to evaluate the frames of each phoneme and the +ve output from each binary classifier taken to form a vector of 26 features that represents the probability of the presence of each specific attribute in the current frame. A separate validation set containing samples from the current phoneme as well as

samples from other phonemes was used to tune the parameters of the OCSVM to achieve the lowest frame-level false-acceptance (FA) and false-rejection (FR) rates. 30% of the validation set was selected from the same phoneme representing the (+ve) samples and 70% randomly selected from the frames of the other phonemes representing the (-ve) ones. We tried three kernels (linear, sigmoid and rbf) with different parameters. Because of the imbalance between the (+ve) and (-ve) samples in the validation set, the optimal parameters were selected to maximize the *F*1 score instead of the overall accuracy to consider both the *FA* and *FR* rates. The *F*1 score is defined as follow:

$$F1 = \frac{2TA}{(2TA + FA + FR)} \quad (1)$$

where *TA* is the true-acceptance.

In the testing mode, all the frames of each phoneme being tested were evaluated and the phoneme acceptance/rejection decision made based on the ratio between the number of in-class and out-of-class frames.

2.4. Goodness of Pronunciation (GOP)

For comparison, we implemented the GOP as proposed in [8] where the posterior probability of each phoneme was estimated using the following equation:

$$P(p_i/O) = \frac{P(O/p_i)}{max_{p_j \in Q}(P(O/p_j))} \quad (2)$$

where p_i is the underlying phoneme and the numerator $P(O/p_j)$ is the phoneme likelihood computed from the forced alignment step and O is the observation segment of p_i obtained from the forced alignment. A free-phoneme recognition step is performed using a phoneme loop grammar created from the list of phonemes in Q. The denominator is the maximum likelihood from the free-phoneme recognition of the observation segment O. The acoustic model used to estimate the GOP is based on the DNN-HMM approach [32].

The normalized log value of the computed score was compared to a predefined threshold to accept or reject the pronunciation. Specific thresholds for each phoneme were tuned to maximize the phoneme *F*1 score in the validation set.

2.5. Speech corpus

The first corpus is the standard TIMIT corpus which consists of recordings of ten phonetically-rich sentences from 630 native-English speakers from 8 different dialects [33]. 462 speakers were used for training while the other 168 speakers were split equally between the validation and testing sets. The training part of this corpus was used to feed the speech attribute detectors, OCSVM models and the DNN-HMM acoustic models, while the validation and testing parts was used for parameter tuning and performance evaluation of each module separately.

Two more corpora were used to evaluate the whole system. The first one consisted of 30 typically developing (TD) children between the ages 6-12 years selected from the OGI kids' speech corpus [34]. Each child pronounced 205 isolated words and 100 short sentences. As the TD dataset is correctly pronounced, we manipulated its phonetic transcription to generate artificial pronunciation errors. We simulated typical CAS substitution errors by altering the phonetic transcription to reflect common substitutions made by children with CAS [5]. The second corpus was recorded from 11 children with CAS producing 450 isolated words. The CAS corpus was collected and annotated by a speech and language pathologist at the University of Sydney.

3. Results

3.1. Speech attribute detection

For each attribute, we trained a binary DNN classifier on the TIMIT corpus to detect the attribute's existence or absence in the current frame. Figure 2 shows the frame-level error rate of the 26 attributes when using shallow NN (1 layer) and deep NN (6 layers). The results show that using 6 layers improved the performance for almost all the attribute classifiers compared to a single layer. The "silence" detector achieved the lowest error rate of around 2% followed by the "affricate", "nasal" and "fricative" with error rates of approximately 5% while the "tense" classifier gave the highest error rate of 15%.

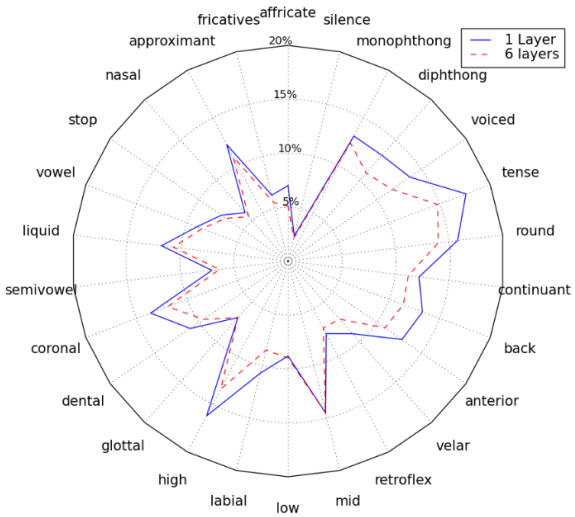

Figure 2: *The error rate of the speech attribute detectors when using 1 layer and 6 layers*

3.2. Phoneme-specific OCSVM model

In this experiment, we used the trained speech attribute classifiers to extract the attribute features from each frame and fed them to the phoneme OCSVM model using the TIMIT corpus. As aforementioned, frames from each specific phoneme were used to train the OCSVM model while only 30% of the frames from the underlying phoneme were included in the validation set and the rest were selected randomly from the other phonemes. The model of each phoneme was trained and tuned separately using the training and validation sets. Overall, the linear kernel gave the worst score for all phonemes followed by the sigmoid while the best performance was obtained using the rbf kernel and achieved *F*1 score greater than 0.8 for the majority of the phonemes. Using the optimum parameters obtained from tuning each phoneme OCSVM model, we tested each model with the separate TIMIT test dataset for a phoneme-level evaluation. The model of each phoneme was tested against samples from the same phoneme to estimate the model *FR* rate and against samples from other phones to estimate the *FA* rate. All samples extracted from the test set were force aligned to the

correct phoneme sequence of each sentence along with its corresponding speech signal.

Table 2 demonstrates the phoneme-level *FR* and *FA* rates for each phoneme and the number of occurrences in the test set. In this experiment, the phoneme is considered in-class (accepted) if the ratio between the in-class frames to the out-of-class frames is greater than 1 otherwise it is rejected which means that the decision threshold is equal to 1. As shown in the table, most of the phonemes had both *FA* and *FR* rates less than 10% with some extremes such as /ih/, /uh/, /eh/ and /ah/.

Overall, consonants perform better than vowels with averages of 5.5 and 5.8 and standard deviations of 1.5 and 2 for the *FA* and *FR* rates respectively compared to averages of 7.2 and 8 and standard deviations of 2.3 and 3 for the vowels *FA* and *FR* rates respectively. The majority of the phonemes have *FA* and *FR* rates lie around 5% while the affricate /ch/ shows the best discriminative performance with both *FA* and *FR* of around 2% and 3% respectively. The *FA* and *FR* rates can be further controlled by relaxing or restricting the decision threshold as will be demonstrated in the next experiment.

Table 2: *The phoneme-level false-acceptance (FA) and false-rejection (FR) rates of the OCSVM model*

Ph	N#	FA (%)	FR (%)	Ph	N#	FA (%)	FR (%)
aa	588	4.93	6.59	g	367	5.99	4.55
ae	743	5.25	5.79	hh	177	4.52	4.22
ah	436	7.8	11.29	jh	180	5	3.54
ao	617	7.29	5.01	k	822	4.62	2.63
aw	118	9.32	8.97	l	1126	6.13	7.64
ay	433	6	4.46	m	644	5.75	5.09
eh	732	11.07	9.12	n	977	6.45	5.65
er	401	5.24	6.99	ng	179	5.03	6.18
ey	395	6.84	6.32	p	456	5.92	4.33
ih	824	7.52	16.6	r	1174	4.94	5.08
iy	1378	5.95	7.79	s	1397	4.01	4.85
ow	413	7.26	8.17	sh	402	7.71	6.7
oy	131	6.11	5.36	t	755	6.49	7.66
uh	105	13.33	12.29	th	131	5.34	10.98
uw	95	5.26	6.18	v	295	6.78	9.64
b	360	4.72	4.99	w	595	2.86	4.39
ch	146	2.05	3.31	y	260	6.92	5.65
d	441	9.52	8.89	z	638	4.86	6.02
dh	279	5.73	5.73	zh	38	7.89	8.53
f	478	4.18	3.54				

3.3. Pronunciation error detection

In this experiment, we used both the TD and CAS speech corpora to show the effectiveness of the algorithm as a pronunciation verification method. Moreover, we compared our OCSVM approach with the GOP as explained in section 2.4. We first force aligned the speech signal with the manipulated version of the phonetic transcription for the TD corpus and with the expected phoneme sequence of the prompt word for the CAS corpus. Forced alignment was performed using DNN-HMM acoustic models trained on the TIMIT training set. Both the OCSVM and GOP methods were then applied to each phoneme and the phoneme accepted if the score exceeded a predefined decision threshold, otherwise rejected. The score of the OCSVM is the ratio between the number of in-class and out-of-class frames while the score of GOP is the normalized estimated log posterior probability. Furthermore, we tuned a phoneme-specific decision threshold for both algorithms to achieve a maximum *F*1 score for each phoneme.

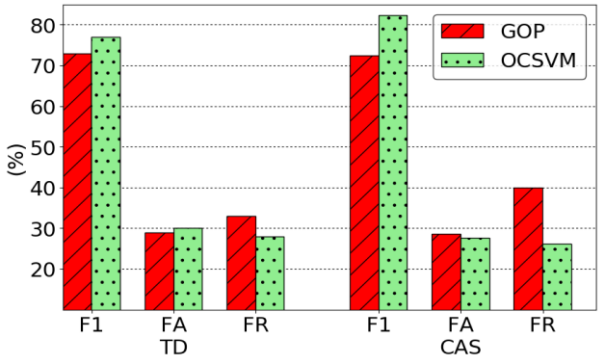

Figure 3: *The F1 scores and the false acceptance (FA) and false rejection (FR) rates of the OCSVM and GOP algorithms applied to the TD test set and the CAS corpus.*

Figure 3 shows the performance of the two algorithms against the TD data set with artificial errors and the CAS data set. The results show that our method slightly outperformed the GOP in the artificial error task with *F*1 score and *FA* and *FR* rates of 0.77, 30% and 28% respectively compared to 0.73, 29% and 33% obtained from the GOP algorithm. On the other hand, our approach showed a significant improvement over the conventional GOP algorithm when applied to data with real pronunciation errors. Our method had a significantly higher *F*1 score of 0.83 and lower *FR* of around 26% compared to the GOP 0.72 and 40%, respectively while both methods achieved similar FA rates.

4. Conclusions

In this work, we presented a phoneme-level pronunciation verification method that leverages upon the anomaly detection framework by using a One-Class SVM model trained with a set of speech attribute features. We first constructed a bank of manner and place of articulations DNN detectors to extract the speech attribute features of each speech frame. A OCSVM model was then trained to learn the distribution of the attribute features of each phoneme from the correctly pronounced data and this model then used to measure the similarity of the new data and decide if the phoneme pronunciation was normal (correct) or anomalous (incorrect).

We compared our method to the DNN GOP algorithm to compare how effective our algorithm is. Both algorithms were applied to the TIMIT corpus with artificial errors and foreign-accented speech corpus. The results showed that our OCSVM method reduced the *FR* rate from 40% when using the GOP method to around 26%.

The results are promising given the limited amount of training data in the TIMIT corpus. Further improvement can be obtained by increasing the data specifically in the attribute detectors as in [24]. Moreover, careful analysis of the attribute features and selection of the best discriminative set of features for each phoneme could further improve the OCSVM model accuracy.

5. Acknowledgements

This work was made possible by NPRP grant #[8-293-2-124] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

6. References

[1] American Speech Language Hearing Association, "Childhood apraxia of speech," 2007.

[2] S. M. Abdou, S. E. Hamid, M. Rashwan, A. Samir, M. Shahin, and W. Nazih, "Computer aided pronunciation learning system using speech recognition techniques."

[3] W. R. R. Dueñas, C. Vaquero, O. Saz, and E. Lleida, "Speech technology applied to children with speech disorders," in *4th Kuala Lumpur International Conference on Biomedical Engineering 2008*, 2008, pp. 247-250.

[4] A. M. Harrison, W. K. Lo, X. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training."

[5] M. A. Shahin, B. Ahmed, J. McKechnie, K. J. Ballard, and R. Gutierrez Osuna, "A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," 2014.

[6] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 193-207, 2017.

[7] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[8] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, pp. 95-108, 2000.

[9] A. Al Hindi, M. Alsulaiman, G. Muhammad, and S. Al-Kahtani, "Automatic pronunciation error detection of nonnative Arabic Speech," in *Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on*, 2014, pp. 190-197.

[10] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, "The goodness of pronunciation algorithm applied to disordered speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[11] B. Mak, M. Siu, M. Ng, Y. C. Tam, Y. C. Chan, K. W. Chan, *et al.*, "PLASER: pronunciation learning via automatic speech recognition," in *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, 2003, pp. 23-29.

[12] O. Saz, S. C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodríguez, "Tools and technologies for computer-aided speech and language therapy," *Speech Communication*, vol. 51, pp. 948-967, 2009.

[13] H. Franco, L. Ferrer, and H. Bratt, "Adaptive and discriminative modeling for improved mispronunciation detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 7709-7713.

[14] T. Zhao, A. Hoshino, M. Suzuki, N. Minematsu, and K. Hirose, "Automatic Chinese pronunciation error detection using SVM trained with structural features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 2012, pp. 473-478.

[15] J. Van Doremalen, C. Cucchiarini, and H. Strik, "Automatic detection of vowel pronunciation errors using multiple information sources," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, 2009, pp. 580-585.

[16] S. Wei, G. Hu, Y. Hu, and R. H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, pp. 896-905, 2009.

[17] K. Truong, A. Neri, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach," in *InSTIL/ICALL Symposium 2004*, 2004.

[18] P. Bonaventura, P. Howarth, and W. Menzel, "Phonetic annotation of a non-native speech corpus," in *Proceedings International Workshop on Integrating Speech Technology in the (Language) Learning and Assistive Interface, InStil*, 2000, pp. 10-17.

[19] B. C. McNeill, G. T. Gillon, and B. Dodd, "Phonological awareness and early reading development in childhood apraxia of speech (CAS)," *International journal of language & communication disorders*, vol. 44, pp. 175-192, 2009.

[20] C. H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101, pp. 1089-1115, 2013.

[21] S. M. Siniscalchi, J. Reed, T. Svendsen, and C. H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, pp. 209-227, 2013.

[22] I. F. Chen, S. M. Siniscalchi, and C. H. Lee, "Attribute based lattice rescoring in spontaneous speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 3325-3329.

[23] S. M. Siniscalchi, D. C. Lyu, T. Svendsen, and C. H. Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data," *IEEE transactions on audio, speech, and language processing*, vol. 20, pp. 875-887, 2012.

[24] D. Yu, S. M. Siniscalchi, L. Deng, and C. H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4169-4172.

[25] J. Li, Y. Tsao, and C. H. Lee, "A study on knowledge source integration for candidate rescoring in automatic speech recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, pp. I/837-I/840 Vol. 1.

[26] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Advances in neural information processing systems*, 2000, pp. 582-588.

[27] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, p. 15, 2009.

[28] S. O. Sadjadi, S. M. Ahadi, and O. Hazrati, "Unsupervised speech/music classification using one-class support vector machines," in *Information, Communications & Signal Processing, 2007 6th International Conference on*, 2007, pp. 1-5.

[29] A. Rabaoui, H. Kadri, Z. Lachiri, and N. Ellouze, "One-class SVMs challenges in audio detection and classification applications," *EURASIP Journal on Advances in Signal Processing*, 2008.

[30] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Proc. Interspeech*, 2015, pp. 2067-2071.

[31] M. Amer, M. Goldstein, and S. Abdennadher, "Enhancing one-class support vector machines for unsupervised anomaly detection," in *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, 2013, pp. 8-15.

[32] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 14-22, 2012.

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic data consortium*, vol. 10, p. 0, 1993.

[34] K. Shobaki, J. P. Hosom, and R. A. Cole, "The OGI kids' speech corpus and recognizers," in *Sixth International Conference on Spoken Language Processing*, 2000.