Speech Modulation Features for Robust Nonnative Speech Accent Detection

Sethserey Sam[1,2], Xiong Xiao[3,5], Laurent Besacier[1], Eric Castelli[2], Haizhou Li[4,5], Eng Siong Chng[3,5]

[1]LIG Laboratory, UMR CNRS 5524 BP 53, 38041, Grenoble Cedex 9, France
[2]MICA research center, UMI CNRS 2954, HUT, Hanoi, Vietnam
[3]School of Computer Engineering, Nanyang Technological University, Singapore
[4]Department of Human Language Technology, Institute for Infocomm Research, Singapore
[5]Temasek Lab@NTU, Nanyang Technological University, Singapore

{sethserey.sam, laurent.besacier}@imag.fr, {xiaoxiong, aseschng}@ntu.edu.sg,
eric.castelli@mica.edu.vn , hli@i2r.a-star.edu.sg

Abstract

In this paper, we propose to use speech modulation features for robust nonnative accent detection. Modulation spectrum carries long term temporal information of speech and may discriminate accents of native and nonnative speakers. For each speech segment to be tested, we extract a 10 dimension feature vector from modulation spectrum and use it for model training and testing. The proposed modulation features are compared with other popular features such as pitch and formant on a nonnative French accent detection task. Results show that the modulation features produce good detection performance and are quite robust to channel distortions. In addition, when combine test scores of modulation features and pitch features, performance is further significantly reduced. The best equal error rate is 13.1% by fusing pitch and modulation-based systems.

Index Terms: nonnative accent, channel mismatch, pitch, formant, prosody, modulation.

1. Introduction

Nonnative accent detection is useful for many speech processing systems. For example, automatic speech recognition (ASR) performance of nonnative speakers can be improved if the model is adapted to the nonnative speakers [1]. However, the adaptation also degrades the performance on native speakers. Hence, it might be good to first detect the nonnative speakers and apply adaptation when it is necessary.

The major challenge in nonnative accent detection is that compared to other factors affecting speech, accent is a weak factor. For example, speech variation due to speaker/gender and channel may be more obvious than speech variation due to accent. To learn the systematic differences between accents, we need speech data from a large amount of speakers. And if we consider channel mismatch between training and testing data, we need data recorded in various channels also; otherwise we may be classifying channels rather than accents. Joint factor analysis [5] is a good way to learn the speaker, channel and accent factors if given enough data for each speaker, channel, and accent. However, in practice it is quite difficult to find such a big database. Hence, we focus on finding feature that is inherently robust to channel mismatch.

Various features have been proposed for nonnative accent detection or classification in the literature. In [2], speech features such as pitch, formant, and frame energy are used for classification of foreign accents in US English. Hidden Markov models (HMM) are used to model the features. In [8], syllable duration, syllable energy, and MFCC were used with HMM modeling. Recently, the voice onset time of three unvoiced stop phones /p/, /t/, and /k/ was proposed for foreign accent classification [12]. Another recent trend is to borrow features from speaker and language recognition systems. For example, phone/word N-gram and MLLR transforms are used as features and support vector machines (SVM) is used as the classifier [13].

In this paper, we propose to use speech modulation spectrum as feature for nonnative accent detection. Speech modulation spectrum [10] captures long term speech dynamics. It has been applied to various speech processing applications, such as speech recognition [11] and speaker recognition [4]. In this study, we extract low dimensional features from high dimensional modulation spectrum representation of speech. The modulation features are compared with several popular features including pitch, formant, MFCC, and phone N-gram.

The rest of the paper is organized as follows. In section 2, we describe the extraction of modulation features and also briefly introduce other features used in this paper. In section 3, we introduce the database, task and evaluation methods, followed by experimental results and discussions in section 4. In section 5, we summarize our findings.

2. Speech Features

In this section, we will describe the extraction of modulation spectrum features and also briefly describe other features.

2.1. Modulation Spectrum Features

The temporal information of speech signal carries information useful for nonnative accent detection. For example, if a nonnative speaker speaks less fluently than native speakers, his/her speaking rate may be slower than native speakers. This difference will help to discriminate the nonnative speaker from native speakers, similar to the use of syllable length or word length as features for nonnative accent detection [8]. However, temporal information not only includes speaking rates, but also the whole speech dynamics along time which may be useful for nonnative accent detection. One way to capture the temporal information is to use the modulation spectrum representation of speech [10]. In this section, we will describe how to compute modulation spectrum of speech signal and how to extract features from the modulation spectrum for nonnative accent detection.

The modulation spectrum used in this study is illustrated in Fig. 1. The left panel of the figure shows the log Mel filterbank coefficients of an utterance. Each filterbank coefficient represents the variation of speech energy along time in the corresponding frequency range. The trajectory of each filterbank is treated as a time sequence and its PSD function is estimated by using fast Fourier transform (FFT). Before FFT, a

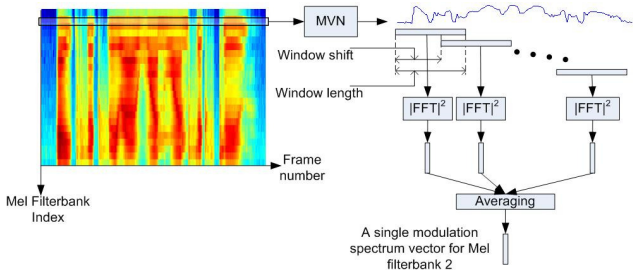

Figure 1: *Generation of modulation spectrum. The second filterbank is used as an example.*

mean and variance normalization (MVN) is applied to normalize the means of the trajectories to zero and variances to one. Besides, as utterances have different lengths, we can apply FFT on fixed-length windowed filterbank coefficients as shown in the figure. This is similar to short-time Fourier transform used to produce spectrogram. Based on our initial study, the window length is set to 0.5s (i.e. 50 frames if frame rate is 100) and the window shift is set to 0.25s. After the FFT, we have a sequence of modulation spectra for each filterbank. As we want to use a fixed-length vector to represent an utterance, we can take the means of the modulation spectra. Hence, there is one average modulation spectrum for each filterbank. Features are extracted from the average modulation spectrum for nonnative accent detection.

Our study shows that the modulation spectra of the filterbanks are highly correlated, which means a redundancy of information. In addition, the dimensionality of modulation spectrum could be high. In our study, we use 23 filterbanks and 64 point FFT (we only retain the first 32 points as Fourier transform coefficients are symmetric), then the dimensionality of the modulation spectrum is 23*32=736. To reduce the dimensionality and de-correlate the features, we apply principal component analysis (PCA) on the modulation spectrum. Only the top 10 projected dimensions with highest variances are used for nonnative accent detection.

Another issue of modulation spectrum is that not all modulation frequencies are useful for our task. The modulation frequency refers to the frequency in the modulation domain. If we generate 100 filterbank frames per second, the modulation frequency has a range of 0-50Hz according to the Shannon theory. Researchers showed that the range 1-16 Hz of modulation frequency is most important for speech recognition while the very low (0-1Hz) and high modulation frequencies (>16Hz) are mainly due to noises [5]. The reason is that human's articulation instruments are constrained and cannot move very fast, hence are not able to generate very high modulation frequency. It is also reasonable to expect that the useful modulation frequencies for nonnative accent detection will be those low modulation frequencies. Hence, it may be necessary to use only selected modulation spectrum bins as input of the PCA. In our study, we found that the first 12 of the 32 modulation frequency bins (corresponding to 0-18.75Hz) carry useful information for our task.

2.2. Pitch

Pitch information is known to discriminate speech accents [2] and has been used in several nonnative accent detection systems [1]. In addition, pitch estimation is relatively robust against noise and channel distortions. In this study, for each utterance, the pitch values of every frame are estimated by using Praat software [3]. The silence and unvoiced frames without valid pitch estimation are removed. As different speakers have different pitch value distribution, e.g. the pitch mean of male speakers and female speakers can be quite different, we apply mean variance normalization (MVN) on the pitch contours (excluding the frames without a valid pitch estimation). Then the delta and acceleration features of the pitch value sequence are computed in the same way as the dynamic features of MFCC. The delta and acceleration measures the slope of the pitch changes which is shown to discriminate accents [2]. As a result, for each voiced frame, there is a 3-dimension pitch feature vector, including the pitch value itself and its delta and acceleration.

2.3. Formant

Formant is also useful for nonnative accent identification [1]. The first two formants (F1 and F2) are extracted by Praat [3] and used as features. The dynamics of formants, i.e. the delta and acceleration, are also computed in the same way as pitch. The dimensionality of formant feature vector is, hence, 6.

2.4. MFCC

MFCC has previously been used for non native accent detection [8]. In this study, 36 dimension MFCC feature vectors are used, including 12 static features (no energy feature), and their delta and acceleration. Silence frames are identified by simple voice activity detection and removed. MFCC features are known to be sensitive to noise and channel distortions. To improve the robustness of MFCC features, we apply the segment-based histogram equalization (HEQ) [9] on all the 36 dimensions independently. The segment length is empirically set to 3 seconds, i.e. HEQ is applied for every 3 seconds of speech.

2.5. PR-VSM

A phonotactic approach PR-VSM (phone recognizer followed by vector space modeling) has been proposed in language recognition [14]. The phone strings, which are the hypotheses of phone recognizer, are converted to a bigram document vector and used as input of SVM. In PR-VSM, a monolingual or multilingual phone recognizer could be used, but according to our experience, the last one gives better performance. In our case, we decode the native and nonnative speech with the multilingual phone recognizer more precisely detailed in our previous study [15]. This multilingual recognizer includes phone models of 3 different languages (English, French and Vietnamese). The motivation is that native and nonnative speakers may use different pronunciation strategies.

3. Experiments

3.1. Corpus and Experimental Framework

The features described in the previous section are evaluated on a nonnative French accent detection task. Three databases are used in the task. The first database (we call it DB1, [6]) contains 3 native French speakers and 16 nonnative French speakers, including 8 Chinese and 8 Vietnamese (In [6], there are only 7 Chinese mentioned, but the 8th Chinese was just added later). There are 100-200 utterances per speaker. The second database (DB2, [16]) contains 12 native French speakers and was recorded in a different channel from DB1. There are 200 utterances per speaker. The third database is the French speech data extracted from the MICA meeting database (see more details in [15]). There are 3 native French speakers and 6 nonnative French speakers: 1 English, 1 Cambodian, and 4 Vietnamese. The MICA data were recorded in different channel from the DB1 and DB2. Hence there is

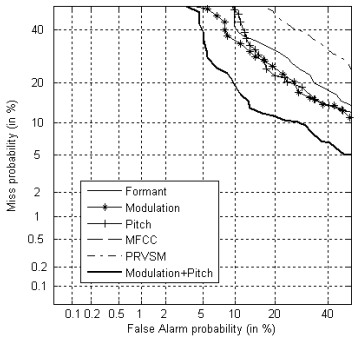

Figure 2: *DET curves on MICA data.*

channel mismatch between any two of the three databases.

We conduct two tasks in our study. In task 1, we use the DB1 and DB2 for model training and testing. There are totally 15 native and 16 nonnative French speakers in DB1 and DB2. A 15-fold cross-validation (CV) is used to produce robust results from the limited database. In each fold, a native and a nonnative speaker are used as test speakers and the rest speakers are used to train models (the last fold contains 2 non native speakers). This design allows all speakers to have a chance to be used as test speakers.

All features are modeled by Gaussian mixture model (GMM) except for the PR-VSM system. Two GMMs are used in each system, one for native speech and the other for nonnative speech. The model complexity of the systems is empirically determined. The numbers of mixtures per class are 4, 64, 64, and 1024 for modulation, pitch, formant, and MFCC features, respectively. Diagonal covariance matrix is used except for modulation features.

One important issue in our study is the channel mismatch between the data sets. As the majority of the native speakers used in the CV (12 speakers from DB2) are recorded in different channel from the other 3 native speakers (3 from DB1) and all the nonnative speakers, there is a danger that the model will learn to classify the channels rather than accents. To evaluate the robustness of features against channel mismatch, we conduct another task, called task 2, in which we test the model trained from DB1 and DB2 on DB3. If a feature happens to classify channel instead of accent in task 1, it should fail in task 2 as the channel in DB3 is different from those of DB1 and DB2. We would like to find robust features which perform well in both task 1 and task 2.

3.2. Evaluation Methods

The task studied here is considered as a detection problem. Our objective is to detect whether a given test utterance is spoken by nonnative French speakers. For each test sentence, a test score is computed. If the test score of a sentence is larger than a pre-defined threshold, the sentence is accepted as spoken by a nonnative speaker. Otherwise, the sentence is rejected as spoken by a nonnative speaker. There are two kinds of errors: false alarm, i.e. a native sentence is detected as nonnative; and missing, i.e. a nonnative sentence is not detected as nonnative. By adjusting the threshold, we can tradeoff the two kinds of errors. Detection error tradeoff (DET) curves [7] will be plotted to show the two kinds of errors with different thresholds. Another evaluation measure is the equal error rate (EER), i.e. the case when the missing rate is equal to the false alarm rate.

For GMM-based systems, the following log likelihood ratio (LLR) is computed and used as test score:

$$LLR(i) = log(p(O_i|GMM_{NN}) / p(O_i|GMM_N)) \quad (1)$$

where O_i is the feature vectors of the i^{th} test utterance, $p(O_i|GMM_{NN})$ and $p(O_i|GMM_N)$ are the likelihood of O_i on the nonnative GMM and native GMM, respectively.

For PR-VSM system, a support vector machine (SVM) with linear kernel [17] is trained to discriminate between native and nonnative conversation sides. The nonnativeness score is the signed distance of the test feature vector from the decision hyperplane (equation 2).

$$f(D_i) = a^T \psi(D_i) + b \quad (2)$$

where D_i is the document vector of the i^{th} test utterance, and $f(D_i)$ represents the signed distance between D_i and the decision surface $a^T \psi(D_i) + b = 0$.

3.3. System Fusion

We also fuse the individual systems to examine whether they are complementary to each other. The scores of individual systems are combined to generate new scores. For each system, the test scores are collected, including both scores of native and nonnative test utterances. As the dynamic range of scores in different systems may be quite different, the variances of the scores are normalized to one by dividing the scores by their standard deviation. After variance normalization, for each test utterances, the fused score is computed as the mean of the individual system scores. Although this fusion scheme is quite simple and may not fully exploit the complementary information of different systems, it is considered to be sufficient for our study here.

4. Results

Let's first examine the EER obtained by the 5 individual systems on both DB1/2 and MICA (DB3) as shown in Table 1. From the table, performance on MICA data is always worse than that on DB1/2 due to channel mismatch. Among the features, pitch features are the most robust against channel distortion with the smallest degradation in EER. This is reasonable as pitch estimation is expected to be almost not affected by channel distortions. Formant features are also quite robust. The modulation features are less robust with the EER tripled. However, it still produces EER of 22.7%, which is the second best among the 5 systems. MFCC and PR-VSM are the least robust, possible due to the high sensitivity of MFCC features to channel distortions (PR-VSM features also rely on MFCC indirectly). In addition, the fusion of the best two features produces even better results as shown in the table (Table 1).

The DET curves of individual systems on MICA data are plotted in Fig. 2. From the figure, pitch and modulation features have the best performance. In addition, the fusion of the two features produces even better results as shown in the figure. In fact, the fusion of pitch and modulation features produces the best performance among all possible combinations. Besides, formant features are also complementary to pitch features and modulation features (not shown in the figure). However, when combining formant, pitch, and modulation together, we get slightly worse performance than only combining pitch and modulation. This may be due to the limitation of our fusion scheme which assigns the same weight to individual systems.

We now investigate the normalized scores in Fig. 3. Each scores' variances are all normalized to 1 for better comparison. The upper panel of the figure shows the scores of native test utterances in the MICA database, each point in the curves represents one test utterance, while the lower panel of the figure shows the scores of six nonnative speakers. The threshold of each detection system (the axis of native and non native decision) is exactly the point where the EER is

recovered (the point where the rate of miss detection is equal to the rate of false alarms). These thresholds are different from one system to another (for example, the threshold of the detection system PR-VSM is 0.3 while that of the MFCC is -0.8). From the figure, it is observed that all detection systems based on MFCC and PR-VSM do not differentiate very clearly native and nonnative speech. Indeed, these two detectors are not robust to the distortions between the training data (DB1 and DB2) and the test data (DB3). Unlike MFCC and PR-VSM features, the detection systems based on the modulation spectrum, the formants and the pitch are more effective for native speech except the speaker "Genevieve" (French female speaker). One possible explanation is that there are very few female native speakers in the training corpus of the detection systems. If we observe on the basis of training data (DB1 and DB2), among the 16 nonnative speakers, there are eight female speakers and eight male speakers. Instead, there are only four French female speakers among the 15 French native speakers. However, the fusion system (Pitch + Modulation) classifies well enough for that speaker (Genevieve). The error comes mainly from the nonnative speakers. For example, we see that the scores of the speaker "Andrew" (native English speakers) and speaker "Sethserey (native Khmer speaker) are almost all misclassified by all systems. This may be due to the fact that their mother tongues (English and Khmer: the official language of Cambodia) are not present in the training data (we have only Vietnamese and Chinese speakers in our training data). Other nonnative speakers are all Vietnamese and the detection performance is good on this subgroup.

Table 1. *Performance of different detection systems*.

Features	EER - DB1/2	EER - MICA
MFCC	13.9 %	51.3 %
Pitch	17.8 %	22.4 %
Formant	19.5 %	26.2 %
PR-VSM	7.9 %	35.1 %
Modulation	7.7 %	22.7 %
Fusion[1]	5.1 %	13.1 %

5. Conclusions

In this paper, we proposed to use modulation spectrum which captures long term dynamics of speech as features for robust nonnative accent detection. Experimental study on a nonnative French accent detection task shows the effectiveness and robustness of modulation spectrum in differentiating native and foreign accents. In addition, modulation spectrum is shown to be complementary to other features, such as pitch features.

6. References

[1] C. Grover, D.G. Jamieson, and M.B. Dobrovolsky. “Intonation in English, French and German: perception and production”. Language and Speech, 30-3:277-295,1987.

[2] L. Arslan and J. Hansen, “Language Accent Classification in American English,” in Speech Communications, vol. 18, no. 4, 1996, pp. 353–367.

[3] Boersma and Paul. Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341-345, 2001.

[4] Kinnunen, T.; , "Joint Acoustic-Modulation Frequency for Speaker Recognition," in the proceedings of ICASSP 2006, vol.1, pp. 14-19 May 2006.

[1] We fused PR-VSM with Modulation for DB1/2 and Pitch with Modulation for MICA data.

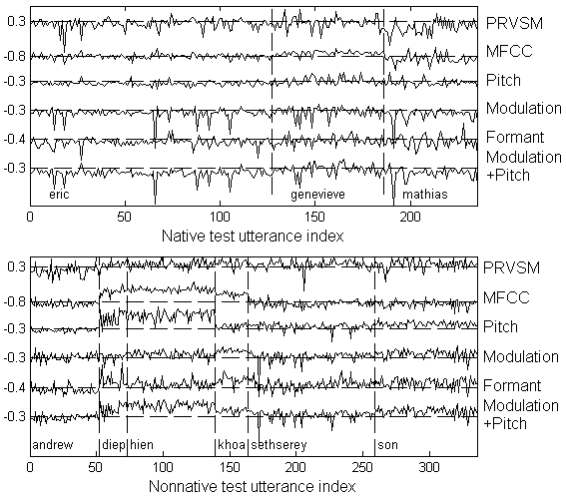

Figure 3: *Scores of test utterances of individual and combined systems. The names of the speakers are also printed just above the x-axis.*

[5] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the relative importance of various components of the modulation spectrum for automatic speech recognition,” Speech Commun., vol. 28, no. 1, pp. 43–55, 1999.

[6] T.-P Tan and L. Besacier, "A French Non-Native Corpus for Automatic Speech Recognition, "LREC 2006, Genoa, pp. 1610-1613, 2006

[7] Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M., The DET curve in assessment of detection task performance. In Proceedings of the European Conference on Speech Communication and Technology, 1997, pp. 1895–1898.

[8] Marina Piat, Dominique Fohr, Irina Illina, Foreign accent identification based on prosodic parameters. In Proceedings of InterSpeech 2008, pp. 759-762.

[9] J. C. Segura, C. Benítez, A. de la Torre, A. J. Rubio, and J. Ramírez, “Cepstral domain segmental nonlinear feature transformations for robust speech recognition,” IEEE Signal Process. Lett., vol. 11, no. 5, pp. 517–520, May 2004.

[10] L. Atlas and S. A. Shamma, “Joint acoustic and modulation frequency,” EURASIP J. Appl. Signal Process., vol. 2003, no. 7, pp. 668–675, 2003.

[11] H. Hermansky, “The modulation spectrum in the automatic recognition of speech,” in Proc. ASRU’97, Santa Barbara, CA, Dec. 1997, pp. 140–147.

[12] John H.L. Hansen, Sharmistha S. Gray, Wooil Kim, “Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification”, Speech Communication, Vol. 52, pp. 777-789, 2010.

[13] E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, M. Akbacak, “Detecting Nonnative Speech Using Speaker Recognition Approaches”, IEEE Odyssey, 2008.

[14] H. Li, B. Ma, C. Lee, “ A vector space modeling approach to spoken language identification”, IEEE Transactions on Audio, Speech, and Language Processing, vol.15(1), pp.271-284, 2007.

[15] S. Sam, E. Castelli, L. Besacier, “Unsupervised acoustic model adaptation for multi-origin non native”, Interspeech 2010. Makuhari, Japan, 2010.

[16] Touch, S., Besacier, L., Castelli, E., & Boitet, C. “Voice aided input for phrase selection using a low level ASR approach - Application to French and Khmer phrasebooks,” SLTU, Penang, Malaysia, 2010.

[17] R. O. Duda, P. E. Hartand, D. G. Stork, “Pattern Classification”, JohnWiley & Sons Inc., 2001.