International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Accent Detection Task to Classify Accented and Non-Accented Speech

Lorita Damayanti, Amalia Zahra

Abstract*: This paper presents a brief survey on accent detection, accent identification, and accent classification. Speech processing has becoming more popular and inspiring expanses lately in signal processing area. It is because speech is one of the most natural form of human communication. However, in processing speech signals intrinsically show many variations even without background noise. Two different person can produce different spectrograms when saying the same sentence. Dialect or Accent is one of the most important factors that can influence the Automatic Speech Recognition or ASR performance besides gender (Unsupervised accent class). Many researches show that dialect or accent in speech can significantly affect the speech system performance. Various methods have been used to increase the accuracy of ASR with accent detection, accent identification, and accent classification. Fused i-vector and Phonotactic are the latest technique that shows a significant degree of accuracy. The purpose of this paper is to briefly survey on accent detection, accent identification, and accent classification and discuss the major improvements made in the past almost 10 years of research.

Keywords* : accent classification, accent detection, accent identification, speech recognition.

I. INTRODUCTION

Speech signal contains information about the person itself, such as the speaker's age, gender, accent, regional, and social background. The first volume of book called Accents of English said that 'accent of English' is defined as "a pattern of pronunciation used by a speaker for whom English is the native language or, more generally by the community or social grouping to which he or she belongs"(Identification of British English regional). Accent or dialect is one of the most important factors that can influence the Automatic Speech Recognition or ASR performance besides gender (Unsupervised accent classification). Accent detection, accent identification , and accent classification have one main privilege that is to have the ability to detect an Accent of a speaker from spoken words and spoken phrase. The difference between them is for accent identification is to identify whether the speaker has a native accent or a non-native accent and for accent classification is to identify whether the speaker has a native accent or a non-native accent and classify them into group components.

Recently, accent identification or AID has rapidly to be of substantial interest in the speech processing community (Identification of British English regional). The ability to recognize accent variations within a language is of importance for forensic speech scientists in speaker comparison applications and speaker profiling (Identification of British English regional). Accent variations are defined by diversities in pronunciation (phone sequence) and speaking style (pitch and rhythm) (Swiss French regional). Accent variations can be divided in two subcategories, the first one is foreign accent and the second one is regional accent (Swiss French regional). The variation of word pronunciation is depending on the speaker native language and the level of foreign language proficiency of the speaker. It gets harder to differentiate and to identify the regional of the speaker when the speaking style changes every word.

Over the past 10 years, many research has been done in the area of accent detection, accent identification, and accent classification. The main purpose of the research, is to reinforce Automatic Speech Recognition or ASR performance. The system which are not influenced by the non-native accent of the speaker, the systems which are adapted to the accent or dialect of the speaker. In most studies, accent identification or AID is used to alleviate this problem because of it's ability to enable the system to use both speech models and pronunciation dictionaries specific to the accent (feature subset selection). In this latter case, in order for the correct set of acoustic models to be chosen AID must occur at recognition time (the impact of accent identif). This approach outperform the GMM-based acoustic methods (contrasting the effect). Several systems has been produced with different acoustic classifiers: a Gaussian Mixture Model-Support Vector Machine (GMM-SVM) system, a Gaussian Mixture Model-Universal Background Model (GMM-UBM) system, GMM-Ngram systems, and a fusion of these methods (iterative classification).

This paper purposely offers the fundamentals of accent detection system along with available various approaches. The accuracy rate of speech recognition can be improved/enhanced by applying one of these techniques in accent detection. This paper also provide results and discussion to acknowledge which approach has succeed improving the accuracy rate.

Manuscript published on 30 September 2019
* Correspondence Author
lorita Damayanti*, student at Bina Nusantara University (Binus), Jakarta, Indonesia.
Amalia Zahra, lecturer at the Master of Information Technology, Binus University, Jakarta, Indonesia.

The rest of the paper is structured as follows. In Section 2 describes the methods that has been used for the past 10 years and the similarity within the methods. Afterwards, we present the results followed by discussion about the methods and the major improvements for the past 10 years in Section 3. The paper ends with conclusions in Section 4.

II. METHODS

Several methods and various features have been proposed for identifying non-native accent or classifying in research area. In (speech modulation), speech modulation spectrum as feature is proposed for non-native accent detection. They extract low dimensional features from high dimensional modulation spectrum representation of speech. Then the modulation spectrum features are compared to pitch, formant, MFCC, and phone N-gram. The result shows that the following features show the effectiveness and robustness to be complementary to other popular features such as pitch features.

In (unsupervised accent class) to improve the performance of Automatic Dialect Identification (DID), various methods has been investigated based on acoustic and text language sub-systems. They propose to use i-Vector system for acoustic approach. To address word selection and grammar factors, a series of Natural Language Processing or NLP techniques are investigated for text language based dialect classification. The NLP techniques that used are two traditional approaches which is N-Gram and Latent Semantic Analysis (LSA). In order to reduce expense, they came up with a cost-effective solution by using web based online podcasts of interviews called UT-Podcast where subjects or speakers speak spontaneously as database. They also consider the NIST LRE-2009 dataset as an addition to UT-Podcast. The result shows that the proposed system shown to improve the Cavg performance by +40.1% and +47.1% on the UT-podcast and LRE-2009 corpora. The audio-text system boosts performance by +6.8% and +4.4% by comparing with the best individual system (i-Vector system).

(Swiss French) is attempting to automatically recognize the speaker's accent among regional Swiss French accents from four different regions of Switzerland: Geneva (GE), Martigny (MA), Neuchatel (NE) and Nyon (NY). This paper propose a generative probabilistic framework for classification based on Gaussian mixture modelling (GMM). There are two GMM-based algorithms are explored: Universal background modelling (UBM) followed by maximum-a-posteriori (MAP) adaptation and total variability modelling (i-Vectors). The database that is used in this paper is a part of PFC database. The results shown that it have validated the first hypothesis of using speaker identification techniques as accent identification. It was shown that the TV-SVM system outperform the GMM baseline which confirms the second hypothesis that i-vector could achieve a higher performance and create a more discriminative feature space. When utterances with more syllables were used, The accuracy of TV-SVM system was slowly increasing.

Another feature is proposed using speech modulation spectrum as feature. It has been applied to various speech processing application such as speech recognition and speaker recognition. In (speech modulation) paper, they used speech modulation as feature and extract low dimensional features from high dimensional modulation spectrum representation of speech. Then the speech modulation features are compared with several features including pitch, formant, MFCC, and phone N-gram. The data that is used in this task are three different databases. The first database labeled as DB1 contains 3 native French speakers and 16 nonnative French speakers, the data includes 8 Chinese and 8 Vietnamese. In DB1, there are 100-200 utterances per speaker. The second database labeled as DB2 contains 12 native French speakers. In DB2, there are 200 utterances per speaker and it was recorded in a different channel from DB1. The third database contains French speech data extracted from the MICA meeting database labeled as DB3. It contains 3 native French speakers and 6 nonnative French speakers. The results show that the modulation features are less robust with the EER tripled with EER of 22.7% which is the second best among 5 systems. The least robust are from MFCC and PR-VSM, likely because of the high sensitivity of MFCC features to channel distortions. The fusion of the best two features produce better results with EER of 5.1% on DB1/2 and EER of 13.1 on MICA.

On the other side, (automatic voice set) has proposed using Voice Onset Time or VOT. VOT is a temporal feature that's defined as the length of time between the release of the oral constriction for plosive production and the onset of vocal fold vibrations (VOT in aphasia). It has proven that VOT can be used to classify mandarin, Turkish, German and American accented English (automatic voice set). They will identify the VOT and the spectral differences and the succeeding vowel of a given stop-vowel sequence. Then the spectral cues are enhanced by one of the four types of feature processing methods such as Discrete Wavelet Transform (DWT), Discrete Mellin Fourier Transform (DMFT) and Discrete Mellin Transform (DMT) using the highest and lowest frequency resolutions (DWTlfr and DWThfr). This paper goal is to develop automatic system to classifies accents using VOT in unvoiced stops. This paper is using CU-Accent corpus as database. The corpus consists of 181 speakers contains 72 males, 107 females and 2 unclassified speakers. The speakers are from a large number of accent groups such as Chinese (Mandarin, Cantonese), Indian (Hindi, Bengali, Marathi, Tamil etc.), native American English, Italian, Spanish, French, Thai, Turkish and many more. The results show that DMT and DWTlfr transformed features ere effective. The average success rate of DMT and DWTlfr are 66.13% and 71.67% respectively. On the other side, the DMFT transformed feature has an average success rate of 79.63%. Table 1 below shows the proposed methods for the past 10 years.

Table 1. Literature Review

Year	**Author**	Method	**Result**
2005	Zheng, et al.	MAP/MLLR	1.0~1.4% absolute reduction of character error rate

2009	Wu, et al.	GMM, LDA, SVM, Feature Subset	Relative improvement of more than 20%
2010	Hansen, et al.	Voice Onset Time (VOT)	VOT is detected with less than 10% error
2011	Sam, et al.	Modulation Spectrum Features	The task shows the effectiveness and robustness of modulation spectrum in differentiating native & foreign accents
2012	DeMarco, et al.	i-Vector method	A new iterative, discrimantive algorithm shown better performance.
2012	Safavi, et al.	GMM-UBM	The SID achieved 100% accurancy and The AID achieved 60.34% accurancy
2013	Bahari, et al.	GMS, GPPS, i-Vector	GPPSs and i-vectors are more effective than GMS
2014	Lazazridis, et al.	UBM-MAP & i-Vector method	TV-SVM outperform GMM baseline
2014	Lazaridis, et al.	SVM, Prosodic features	As the number of the syllables increases the higher accurancy improves (Long utterances)
2014	Lacheret, et al.	Prosodic Prominence	They pointed out how the interplay between manual and automatic data processing is necessary to provice valuable insight into the bias inevitably associated with manual annotation.
2015	Mannepalli, et al.	Prosodic & Formant features	Overall Efficiency obtained is 72%
2015	John H.L. Hansen & Gang Liu	Fused i-Vector & TF-IDF	Improved Cavg performance by +40.1% and +47.1% relatively
2016	Najafian, et al.	Fused i-Vector & Phonotactic	Fusion of i-vector and phonotactic results in a higher accurancy (84.87%)

Invthe study in i-vector was compared to twoesssthe study in twot

III. RESULTS AND DISCUSSION

Accent is one of the crucial problem when it comes to signal processing. The ability to detect a native or a non-native speaker is a challenging task. Several methods and features have been proposed by researchers. The most recent methods/features are improved and updated compared to the old one. As seen on table 1, the latest method using fused I-vector system and Phonotactic shows a high accuracy of 84.87%. This paper (identif of british) compared their results with the research about accent identification before which is a research from DeMarco. et al. Table 2 shows the comparison between Najadian.et al. and DeMarco. et al.

Table 2. Comparison Results between both proposed system (identif of british) [11] [1]

AID systems	Acoustic systems	Phonotactic systems	Final fused AID performance
DeMarco et. al. [43]	AID (630 Ac.) 81.05%	—	Fused (630 Ac.) 81.05%
Our proposed AID Systems	AID (1 Ac.) 76.76%	AID (15 Phon.) 80.65%	Fused (1 Ac. & 15 Phon.) 84.87%

As seen on table 2, DeMarco et. al. has higher accuracy of Acoustic systems compared to Najadian et. al. proposed systems. Figure 3 shows the information about Najadian's system consist of one i-vector system and 15 phonotactic systems. DeMarco's system uses much higher UBM size and The T-matrix rank of a combination of 630 subsystems compared to Najadian's system. As shown on table 2, Najadian's system outperforms DeMarco et al. proposed system by 4.7% using acoustic and phonotactic fused Accent Identification (AID) system. Fig 1. Shows the fusion of SVM scores from 15 paarallel PRLM-SVMs.

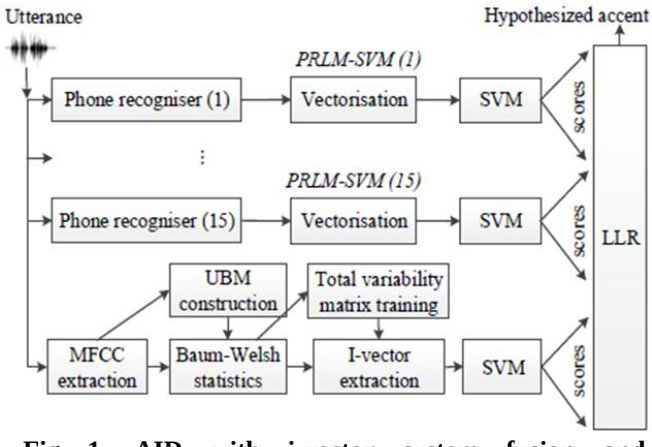

Fig 1. AID with i-vector system fusion and PRLM-SVM (identif of british) intwoesVnumberFrequency is used to recognize the accent based on the use of certain phonetic sequencesprocessing stages. Theyese are

IV. CONCLUSION

A. Figures and Tables

This paper review try to briefly collect the majority of several approaches in Accent detection area. This paper also focusing on the selection of the previous studies and the used techniques in order to know which technique has better performance and is updated and comparing them according on the research results. This paper have presented many techniques and features on improving the performance of Speech recognition using accent detection but clearly the most recent one was better and most used by researchers. Fused i-vector and Phonotactic have proven that these techniques can be on the options for increasing the ASR performance using accent detection.

REFERENCES

1. M. Najafian, S. Safavi, P. Weber and M. Russell, "Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic system," *Odyssey*, pp. 132-139, 2016.
2. J. H. L. Hansen and G. Liu, "Unsupervised accent classification for deep data fusion of acoustic and language information," *Elsevier*, vol. 78, no. April 2016, pp. 19-33, 2015.
3. P. N. Sastry, "Accent Detection of Telugu Speech using Prosodic and Formant Features," pp. 318-322, 2015.
4. D. Yu and L. Deng, Automatic Speech Recognition A Deep Learning Approach, Springer, 2015.
5. Lacheret-dujour, A.-c. Simon, J.-p. Goldman, A. Lacheret-dujour, A.-c. Simon, J.-p. Goldman and M. A. Prominence, "Prominence perception and accent detection in French : from phonetic processing to grammatical analysis," no. 39, pp. 95-106, 2014.
6. H. Kamper and T. Niesler, "The impact of accent identification errors on speech recognition of South African English Accents of English in South Africa," South African Journal of Science, vol. 110, no. 1, pp. 1-6, 2014.
7. Lazaridis, J.-p. Goldman and M. Avanzi, "Syllable-based Regional Swiss French Accent Identification using Prosodic Features," Nouveaux cahiers de linguistique francaise, vol. 1, no. 2000, pp. 1-2, 2014.
8. M. Avanzi and P. N. Garner, "Swiss french regional accent identification," in Odyssey: The Speaker and Language Recognition Workshop, 2014.
9. M. H. Bahari, R. Saedi, H. V. hamme and D. V. Leeuwen, "Accent Recognition using i-Vector, Gaussian Meana Supervector and Gaussian posterior probability Supervector for spontaneous Telephone Speech," IEEE, no. October 2013, pp. 7344-7348, 2013.
10. S. Safavi, A. Hanani, M. Russell, P. Jan and M. J. Carey, "Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identi fi cation," IEEE Signal Processing Letters, vol. 19, no. 12, pp. 829-832, 2012.
11. DeMarco and S. J. Cox, "Iterative Classification of Regional British Accents in i-Vector space," Machine Learning in Speech and Language Processing (MLSLP), no. September 2012, pp. 1-4, 2012.
12. Y. Y. Tan , "Age as a factor in ethnic accent identification in Singapore," Journal of Multilingual and Multicultural Development, vol. 33, no. December 2014, pp. 37-41, 2012.
13. S. Sam, X. Xiao, L. Besacier, E. Castelli, H. Li, E. S. Chng, U. M. R. C. Bp and G. Cedex, "Speech Modulation Features for Robust Nonnative Speech Accent Detection School of Computer Engineering , Nanyang Technological University , Singapore Department of Human Language Technology , Institute for Infocomm Research , Singapore Temasek Lab @ NTU ,," Interspeech, pp. 2417-2420, 2011
14. J. H. L. Hansen, S. S. Gray and W. Kim, "Automatic voice onset time detection for unvoiced stops (/ p /,/ t /,/ k /) with application to accent classification," Speech Communication, vol. 52, no. 10, pp. 777-789, 2010.
15. P. Auzou, C. Ozsancak, R. J. Morris, M. Jan, F. Eustache and D. Hannequin, "Voice onset time in aphasia , apraxia of speech and dysarthria : a review Voice onset time in aphasia , apraxia of speech and dysarthria : a review," Clinical Linguistics & Phonetics, vol. 14, pp. 131-150, 2009.
16. T. Wu, J. Duchateau, J.-p. Martens and D. V. Compernolle, "Feature subset selection for improved native accent identification," Speech Communication, vol. 52, pp. 83-98, 2009.
17. Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr and S.-y. Yoon, "Accent Detection and Speech Recognition for Shanghai-Accented Mandarin," in NTERSPEECH 2005 , Lisbon, 2005.

AUTHORS PROFILE

lorita Damayanti is a student at Bina Nusantara University (Binus), Jakarta, Indonesia. She is currently taking her bachelor and master degree at the same time. She is taking Information Technology major, The Faculty of Computer Science. She attended the training of Assistant Laboratory in 2016. She worked as a Quality Control, Department of Management Information System at PT.MNC Sky Vision as an Internship in 2018. She has never done any research before. This is her first research in the area of speech technology. profile which contains their education details, their publications, research work, membership, achievements, with photo that will be maximum 200-400 words.

Amalia Zahra She obtained her PhD degree from the School of Computer Science and Informatics, University College Dublin (UCD), Ireland in 2014, and her bachelor degree from the Faculty of Computer Science, University of Indonesia in 2008. She does not have a master degree. After finishing her bachelor degree in 2008, she worked as a research assistant for a year, and then continued to pursue PhD immediately afterwards.. The topic of her PhD thesis was about spoken language identification. Her research interest has been in the area of speech technology, machine learning, and computational linguistics since 2008. . Therefore, her publications are related to such topics. Currently, she is working as a lecturer at the Master of Information Technology, Binus University, Jakarta, Indonesia.profile which contains their education details, their publications, research work, membership, achievements, with photo that will be maximum 200-400 words.