Development of Acoustical Feature Based Classifier Using Decision Fusion Technique for Malay Language Disfluencies Classification

Raseeda Hamzah*[1], Nursuriati Jamil[2], Rosniza Roslan[3]
Faculty of Computer and Mathematical Science,
MARA University of Technology, Shah Alam, 40450 Selangor, Malaysia
*Corresponding author, e-mail: rashamzah82@gmailcom[1], liza@tmsk.uitm.edu.my[2],
rosniza@tmsk.uitm.edu.my[3]

Abstract

Speech disfluency such as filled pause (FP) is a hindrance in Automated Speech Recognition as it degrades the accuracy performance. Previous work of FP detection and classification have fused a number of acoustical features as fusion classification is known to improve classification results. This paper presents new decision fusion of two well-established acoustical features that are zero crossing rates (ZCR) and speech envelope (ENV) with eight popular acoustical features for classification of Malay language filled pause (FP) and elongation (ELO). Five hundred ELO and 500 FP are selected from a spontaneous speeches of a parliamentary session and Naïve Bayes classifier is used for the decision fusion classification. The proposed feature fusion produced better classification performance compared to single feature classification with the highest F-measure of 82% for both classes.

***Keywords**: Decision fusion; Naïve Bayes; acoustical feature; Random Forest; disfluencies detection*

1. Introduction

Speech is the most common way of interaction used by humans [1]. Speech can be categorized as read and sponatneous. The daily communication type of speech used by human is spontaneous speech that highly contain filled pause. Filled pause (FP) is vocalized pause and non-lexical speech event that is usually used by speakers to prevent interruption from others while planning their utterances [2]. The importance of FP detection or handling can be viewed in several areas [3]. In automatic speech recognition (ASR), FP detection is viewed as essential [4] as it is recognized as one of ASR performance degradation factor [5]. One of the ways of dealing with FP is by detecting and removing it. However, the main problem of detecting FP is the occurrences of elongation (ELO) which is an extended syllable in a word that has the same acoustical features with FP. Discriminating filled pause against elongation is critical because ELO is semantically meaningful unlike FP. Removing ELO in a speech sentence will change its [4] semantic context.

A number of FP research has been done for various languages including English [6], Mandarin [2], Portuguese [7], Slovenian [8], Hungarian [9] and Polish [10]. Although FP has similar acoustical speech features pattern, it is language dependent [8]. Speech-related research in Malay language is still at an early stage [11] and disfluency detection for Malay language is scarce. Malay language is categorized as an under-resourced language due to its lack of electronic resources for speech and language processing such as monolingual corpora, transcribed speech data and pronunciation dictionaries [12]. Thus, it is empirical that speech-related work in Malay language is been pursued.

One of a well-established FP detection methods is by utilizing the acoustical features of FP [6]. Standard acoustical features such as formant frequency (FF), Mel-frequency cepstral coefficients (MFCC), fundamental frequency (F0), and short time energy (STE) have previously been used in FP researches. Other than the aforementioned acoustical features, popular speech features such as zero crossing rates (ZCR) and speech envelope (ENV) are not being tested thoroughly for FP and ELO classification. ZCR has been used in non-lexical speech event detection but focusing on laughter, applause and cheer [13]; to determine the voice and

Received June 12, 2017; Revised August 29, 2017; Accepted September 13, 2017

unvoiced region [14] and also in vowel and consonant classification [15]. The use of ZCR can hardly be found to discriminate FP and ELO. ENV is also another feature that is utilized in speech processing research. One of its usages is to find the syllable nuclei. ENV is seen as a potential feature in this research with the fact that Malay language is alphabetic-syllabic and the FP and ELO in Malay language is represented syllabically. Thus, ZCR and ENV along with the other eight well-established FP features are used as speech features for classification of Malay language FP and ELO.

Fusion technique for classification introduced in ASR research is known to enhance accuracy compared to single feature [16]. In this research, each of the acoustical features has its own role in classifying speech. Therefore, the advantage of each acoustical feature is utilized in the classification performance. This concept is called decision fusion and also chosen because it is computationally more efficient than the other fusion types [17]. Although there are various fusion techniques available, we implemented the Naïve Bayes assumption to fuse the classifier to get the final decision as it is a simple probabilistic classifier that allows the fusion of different feature and has become one of the famous techniques in speech research [18].

2. Speech Data Collection

The speech data used in this research is gathered from hansard documents of Malaysian Parliament’s debate sessions of 2008 [19]. It comprises Malay language spontaneous speeches spoken by male and female speakers of Malay, Chinese and Indian ethnics. Since the speech data was recorded live, it is surrounded with background noise, interruptions, and various speaking style (low, medium and high intonation).

A total of 800 sentences are selected from the speech data for our experiment. Five hundred (500) FPs and 2000 normal words of different duration and multi-speakers are then gathered from these sentences. These two datasets are defined as FP_dataset and Word_dataset, respectively. We subsequently extracted 500 elongated words from Word_dataset and define it as ELO_dataset. From these datasets, 70% is used as training and 30% is used as testing. The selection of the elongated data is based on the most common uttered words in ELO_dataset.

Malay words are agglutinative alphabetic-syllabic that are based on four distinct syllable structures, i.e. V, VC, CV and CVC [19]. Few examples of ELO are tabulated in Table 1. In English language, ELO is described as the extension at the end of the utterances as a replacement of FP [20]. Based on our data analysis, we described our ELO as the last syllable of an utterance that can be at any location in a sentence.

Table 1. Elongated Malay language word structure

WORD	STRUCTURE
ADA	V+CV
BAHAWA	CV+CV+CV
BERAPA	CV+CV+CV
BILA	CV+CV
JUGA	CV+CV
KATA	CV+CV
KERANA	CV+CV+CV
NYA	CCV
MAKA	CV+CV
TANYA	CV+CCV
NEGARA	CV+CV+CV
SAYA	CV+CV
SECARA	CV+CV+CV
MEREKA	CV+CV+CV

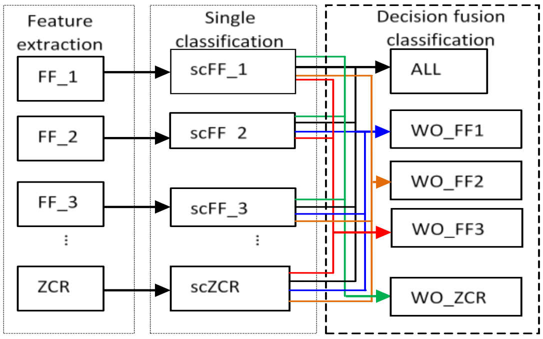
C-Consonat; V-Vowel

3. Feature Extraction

To observe the FP and ELO characteristics, ten acoustical speech features are extracted in this research. Prior to the feature extraction, a standard speech pre-processing such as windowing, framing and pre-emphasising are done. Pre-processing is imporatnt in any signal analysis research to get the important information from the raw data [21]. The first acoustical feature is four levels of formant frequencies (FF1, FF2, FF3 and FF4). The linear prediction coding (LPC) is used in the extraction process [22]. Previously, formant frequency is extracted and evaluated to analyse FP's pattern stability [6], [9]. Then, the fundamental frequency (F0) is extracted by using Average Magnitude Differences Function (AMDF) due to its computation's simplicity and reliability in tracking voice pitch contour [23]. The extraction is then preceded with 12-Mel frequency cepstral coefficient (MFCC) to produce a multi-dimensional feature vector for every frame of speech. The other features are speech envelope by extracting the local maxima of the speech [24]; Zero Crossing Rates (ZCR) is extracted as in [13]; root mean square energy (RMSE) [25].

4. Classification

In the classification stage, two types of Naïve Bayes classifications are performed. The first type is single feature and the second type is multiple features classifications. The whole embodiment of the proposed classification for FP and ELO is illustrated as in Figure 1. The proposed Naïve Bayes classifier utilized 10-fold cross validation as recommended by [26]. The 10-fold validation is implemented to ensure that every instance is evaluated in the classifier and to ensure that the classifier is able to generalize on each datasets. The Kernel density function is used on the training data to get the probability distribution function of the data.



sc - single classification, WO_FF1:WO_ZCR – without feature

Figure 1. Proposed Naïve Bayes classification model for FP and ELO

4.1. Single feature classification

From Bayesian view, the classification problem of this research is to recursively compute the degree of belief in the object class *C*, being a FP/ELO class, given the feature *d*. Based on Bayes theorem in the case of FP and ELO classification as in (1).

$$P(C_j \mid d) = \frac{P(d \mid C_j)p(C_j)}{P(d)} \tag{1}$$

where, C_j = class either FP or ELO and *d* is the acoustical feature vector of tested data.

$P(Cj \mid d)$ =the posterior probability of sample *d* being in class $C_{j,}$

$P(d \mid Cj)$ =the conditional probability of producing sample *d* given class C*j*,

$P(d)$ =the probability of occurrences of sample *d*, and

$P(Cj)$ =the prior probability of FP or ELO.

4.1.1. Single feature classification

An additional analysis on each feature is also taken to rate the feature's importance in representing the FP and ELO. This analysis is done by using Boruta-Random Forest algorithm [27]. In Boruta, the information system is extended with shadow. Shadow is known as artificial features that are created by permuting the order of values in the original data. These artificial features are used to gather the shadows' importance scores to judge the significance of the scores obtained by the actual features. The score is measured using mean Z-score.

4.1.2. Decision Fusion Classification

The decision fusion for multiple features is implemented by using Naïve Bayes. In Naïve Bayes decision fusion, each feature went through a preliminary classification producing a conditional probability $P(d_n | C_j)$. Then, the $P(d_n | C_j)$ of every single classifier is merged by using product rules as in (2).

$$P(d_n | C_j) = P(d_1 | C_j) * P(d_2 | C_j) * ... P(d_n | C_j) \tag{2}$$

Thus Equation (2) can be rewritten as (3) and (4)

$$P(C_j \mid d_n) \propto P(C_j) \prod_{n}^{i} P(d_n \mid C_j) \tag{3}$$

$$\sum_{C_j} P(C_j | d_n) = 1 \tag{4}$$

For example of two features, is the probability of class *Cj* generating the observed value for first acoustical feature d_1, multiplied by the probability of class *Cj* generating the observed value for second acoustical feature d_2. The steps for decision fusion classification are as below:

Step 1: All features are used in the classification. These features are labeled as 'ALL'.

Step 2: FF1 is excluded for classification process leaving 9 features. These features are labeled as without FF1 (WO_FF1).

Step 3: The eliminating process is repeated for each feature, XX, labeling them as WO_XX. XX refers to (FF1: ZCR).

5. Evaluation

This research used three types of measurement to evaluate the classification performance. The measurements are precision, recall and F-measure as in (5), (6) and (7).

$$\mathrm{Re}call = \frac{\#correctly\_\mathrm{det}ected\_FP/ELO}{\#FP/ELO} \tag{5}$$

$$\mathrm{Pr}ecision = \frac{\#correctly\_\mathrm{det}ected\_FP/ELO}{\#\mathrm{det}ected\_FP/ELO} \tag{6}$$

$$F-Measure = (\frac{\mathrm{Re}call \times \mathrm{Pr}ecision}{\mathrm{Re}call + \mathrm{Pr}ecision}) \times 2 \tag{7}$$

6. Result and Discussion

The classification of FP and ELO between single and decision fusion classifier is compared. The importance ranking of each acoustical feature done using Boruta [27] is also presented in Table 2. Results showed that ZCR ranked as the most important scoring at 33 followed by ENV achieving a score of 28. This shows that ZCR has the ability in discriminating FP and ELO compared to other features, because of its capability in detecting vowel and consonant. The FP in this research consists of vowels, whereas ELO comprises consonant and vowel.

Table 2. Results of single feature classification

Feature	Recall%		Precision%		F-Measure%		Boruta
	ELO	FP	ELO	FP	ELO	FP	Z-score
FF1	62 ± 1.77	58 ± 1.95	51 ± 1.88	69 ± 1.33	56 ± 1.52	63 ± **2.20**	12
FF2	60 ± 1.52	66 ± 1.26	57 ± 1.71	69 ± 1.92	63 ± 1.36	62 ± 1.89	22
FF3	62 ± 1.82	63 ± 0.59	49 ± 2.00	79 ± 1.84	55 ± 2.01	70 ± 1.72	14
FF4	50 ± 1.74	46 ± 2.02	32 ± 1.29	45 ± 1.74	48 ± 1.97	37 ± 1.29	01
F0	56 ± 1.99	58 ± 1.62	39 ± 1.13	73 ± 1.77	46 ± 1.88	65 ± 1.94	06
STE	43 ± 2.03	48 ± 1.84	33 ± 0.77	59 ± 1.46	45 ± 1.92	42 ± 0.73	01
RMSE	46 ± 1.39	56 ± 1.99	67 ± 0.69	34 ± 1.59	55 ± 0.94	42 ± 2.11	07
MFCC	55 ± 0.57	56 ± 1.73	40 ± 0.37	76 ± 0.93	46 ± 2.12	64 ± 0.53	04
ENV	74 ± 0.55	80 ± 1.55	60 ± 1.83	81 ± 0.86	66 ± 0.88	**80** ± 1.69	**28**
ZCR	76 ± 1.48	72 ± 1.92	65 ± 1.72	85 ± 1.79	**70** ± 1.69	78 ± 1.51	**33**

Table 3. Results of decision fusion classification

Feature	Recall%		Precision%		F-Measure%	
	ELO	FP	ELO	FP	ELO	FP
ALL	78 ± 1.82	86 ± 1.24	85 ± 2.10	78 ± 0.87	82 ± 2.09	81 ± 1.77
WO_FF1	75 ± 2.03	83 ± 0.99	82 ± 1.53	76 ± 0.59	79 ± 1.97	79 ± 1.64
WO_FF2	83 ± 1.91	78 ± 1.57	77 ± 1.97	84 ± 1.73	80 ± 1.82	80 ± 0.88
WO_FF3	75 ± 1.75	87 ± 1.92	85 ± 1.63	78 ± 1.79	81 ± 0.64	81 ± 0.97
WO_FF4	79 ± 0.84	83 ± 2.11	81 ± 1.89	75 ± 1.89	81 ± 0.48	78 ± 0.57
WO_F0	73 ± 0.97	88 ± 0.79	85 ± 1.73	77 ± 1.87	80 ± 0.75	81 ± 0.35
WO_STE	80 ± 0.72	84 ± 0.94	79 ± 1.92	85 ± 2.01	**82** ± 1.88	**82** ± 1.33
WO_RMSE	81 ± 1.69	82 ± 1.82	79 ± 1.35	83 ± 2.19	81 ± 1.52	81 ± 2.08
WO_MFCC	76 ± 1.88	81 ± 1.93	80 ± 1.59	79 ± 0.79	78 ± 1.87	79 ± 1.74
WO_ENV	70 ± 1.53	78 ± 1.79	76 ± 1.49	73 ± 0.93	74 ± 1.93	74± 1.28
WO_ZCR	73 ± 1.29	75 ± 1.83	71 ± 1.27	76 ± 0.89	74 ± 1.55	73 ± 1.17

The results are represented as mean and standard deviation (mean ± standard deviation), for each 10 times of independent trial (10-fold CV). The highest standard deviation of 2.20 indicates that the result of each fold is consistent.To test the performance of each acoustical feature, the FP and ELO is classified using Bayes classifier and the results for single feature classification are tabulated in Table 2. From Table 2, the highest F-Measure for FP is denoted by ENV feature with 80%. For ELO, the highest F-measure is scored by ZCR with 70%. The result of Boruta's algorithm is consistent with to the single classification's result i.e. the selected features of ZCR and ENV scored the top 2 ranks and when fed into classifier they performed the top two F-Measures.

The findings suggest that these two features are important in representing FP and ELO. The results of decision fusion are tabulated as in Table 3. Table 3 shows that the F-measure significantly improved when the decision fusion is applied on each feature. In the Naïve Bayes decision fusion, classification of FP and ELO in each single classifier complemented each other to amend the final classification. For example, the second test of a dataset cannot be detected correctly with ZCR classifier; however it can be detected with ENV which consequently improves the final result. The choice of dataset (ALL: WO_ZCR) demonstrates the effect of feature relevance towards the Naive Bayes based decision fusion on FP and ELO classification. The average F-measure score is about 80% which only differs at 3% for each dataset for all cases except for WO_ENV and WO_ZCR. However, when the ENV and ZCR is eliminated (i.e. WO_ENV and WO_ZCR), the result significantly decreased by 6% for both FP and ELO, indicating that Naïve Bayes fusion is effective when relevant feature is used.

Acknowledgement

This research was supported by Ministry of Higher Education (MOHE) with grant reference of (600/RMI/DANA 5/BESTARI/0002/2016) and Universiti Teknologi MARA. We thank our colleagues Dr. Norizah Ardi from Akademi Pengajian Bahasa (APB) UiTM and Dr.Noraini Seman from Faculty of Computer Science and Mathematic who provided insight and expertise that greatly assisted the research.

References

[1] Esmaileyan, Z, Marvi, H. A Database for Automatic Persian Speech Emotion Recognition: Collection, Processing and Evaluation. *International Journal of Engineering Transaction A: Basics* 2014; 27: 79-90.

[2] Li, YX, He, QH, Li, T. A novel detection method of filled pause in mandarin spontaneous speech. *Computer and Information Science,* Seventh IEEE/ACIS International Conference. 2008: 217-222.

[3] Garg, G, Ward, N. Detecting Filled Pauses in Tutorial Dialogue. *Tech. report, Univ. Texas El Paso*, 2006: 1–9.

[4] Li, YX, He, QH, Li, W, Wang, ZF. *Two-level approach for detecting non-lexical audio events in spontaneous speech*. Audio Language and Image Processing (ICALIP), International Conference. 2010: 771-777.

[5] Stouten, F, Martens, JP. *A feature-based filled pause detection system for Dutch*. Automatic Speech Recognition and Understanding ASRU'03 IEEE Workshop. 2003: 309-314.

[6] Audhkhasi, K, Kandhway, K, Deshmukh, OD, Verma, A. *Formant-based technique for automatic filled-pause detection in spontaneous spoken English*. Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference. 2009: 4857-4860.

[7] Veiga, A, Candeias, S, Lopes, C, Perdigão, F. *Characterization of hesitations using acoustic models*. Proc. of the 17th International Congress of Phonetic Sciences, ICPhS XVII, (2011), 2054-2057.

[8] Žgank, A, Rotovnik, T, Sepesy Maučec, M. Slovenian spontaneous speech recognition and acoustic modeling of filled pauses and onomatopoeas. *WSEAS Transactions on Signal Processing*, (2008).

[9] Deme, A, Markó, A. Lengthenings and filled pauses in Hungarian adults' and children's speech. *Proceedings of DiSS*. 2013: 21–24.

[10] Karpiński, M. Acoustic Features of Filled Pauses in Polish Task-Oriented Dialogues. *Archives of Acoustics.* 2013; 38(1): 63-73.

[11] Fook, CY, Hariharan, M, Yaacob, S, Adom, A. *A review: Malay speech recognition and audio visual speech recognition*. Biomedical Engineering (ICoBE), IEEE International Conference, (2012), 479-484.

[12] Besacier, L, Barnard, E, Karpov, A, Schultz, T. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 2014: 85-100.

[13] Cai, R, Lu, L, Zhang, HJ, Cai, LH. *Highlight sound effects detection in audio stream*. Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 IEEE International Conference 2003; 3: III-37.

[14] Jalil, M, Butt, FA, Malik, A. *Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals.* Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), IEEE International Conference 2013: 208-212.

[15] Ito, MR, Donaldson, RW. Zero-crossing measurements for analysis and recognition of speech sounds. *Audio and Electroacoustics, IEEE Transactions* 1971; 19(3): 235-242.

[16] Planet, S, Iriondo, I. *Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition*. Information Systems and Technologies (CISTI), 2012 7th Iberian Conference 2012: 1-6. IEEE.

[17] Cremer, F, Schutte, K, Schavemaker, JG, den Breejen, E. (2001). A comparison of decision-level sensor-fusion methods for anti-personnel landmine detection. *Information fusion*, (2001); 2(3): 187-208.

[18] Lee, LW, Low, HM, Mohamed, AR. A Comparative Analysis of Word Structures in Malay and English Children's Stories. *Pertanika Journal of Social Sciences & Humanities*, 2013; 21(1): 67–84.

[19] Seman, N, Bakar, ZA, Bakar, NA. *Measuring the performance of isolated spoken Malay speech recognition using Multi-layer Neural Networks*. Science and Social Research (CSSR), 2010 International Conference 2010: 182-186. IEEE.

[20] Goto, M, Itou, K, Hayamizu, S. A real-time filled pause detection system for spontaneous speech recognition. *Eurospeech 99*. 1999: 227–230.

[21] Hamidi, H, Daraei A. Analysis of Pre-processing and Post-processing Methods and Using Data Mining to Diagnose Heart Diseases. *International Journal of Engineering (IJE), Transactions A: Basics* 2016; 29(7): 921-930.

[22] Eide, A´OC. Linear Prediction. *Report of Dublin Institute of Technology,* (2008).

[23] Ross, MJ, Shaffer, HL, Cohen, A, Freudberg, R, Manley, HJ. Average magnitude difference function pitch extractor. *Acoustics, Speech and Signal Processing, IEEE Transactions.* 1974; 22(5): 353-362.

[24] Reddy, AA, Chennupati, N, Yegnanarayana, B. Syllable nuclei detection using perceptually significant features. *Proc. of Interspeech*. 2013: 963–967.

[25] Sakhnov, K, Verteletskaya, E, Simak, B. *Dynamical energy-based speech/silence detector for speech enhancement applications.* Proceedings of the World Congress on Engineering. 2009; 1: 2.

[26] Paja, W, Wrzesień, M. *Melanoma important features selection using random forest approach*. 6th International Conference on Human System Interactions (HSI), 2013; 11(1), V12.

[27] Kursa, MB, Jankowski, A, Rudnicki, WR. Boruta A System for Feature Selection. *Fundamenta Informaticae,* 2010; 101(4): 271-285