Analyzing and Classifying Indonesian Spontaneous and Dictated Speech

Cil Hardianto Satriawan*, Dessi Puji Lestari†
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia 40132
*23515053@std.stei.itb.ac.id, †dessipuji@stei.itb.ac.id

***Abstract*—The accurate recognition of spontaneous speech is crucial in achieving practical speech recognition. Statistical-based recognition models typically employ a large amount of read or dictated speech for training, which often yields poor spontaneous recognition performance. Many approaches have been forwarded to improve performance, including model adaptation and model switching. In an effort to improve Indonesian language spontaneous recognition performance, we attempt to pinpoint the acoustic differences between spontaneous and dictated Indonesian speech. At the phoneme level, we find that there are differences in the distribution and pronunciation of several key phonemes associated with filled pauses. Across speakers, there is a consistent reduction in segment duration and segment energy, with a less marked spectral reduction. Using these differences as a starting point, we train a number of classifiers that can accurately identify spontaneous and read Indonenesian utterances at F1 scores consistently above 90%. We show that classification is achievable by considering segment features and feature differences between consecutive segments, or "delta" and "delta-delta segments".**

***Index Terms*—Speech recognition, spontaneous speech, Indonesian language**

I. Introduction

The accurate recognition of spontaneous speech is one of the large ongoing challenges towards achieving practical speech recognition, with many systems still performing poorly. This is seemingly in contrast to the recognition performance achievable on most state-of-the-art statistical-based systems on dictated speech tasks and limited domain spoken interactions [15].

This poor performance may be attributed in part to the fact that a large proportion of the data used to train models is derived from dictated instead of spontaneous speech. Recording and labelling large amounts of new spontaneous speech is a difficult and costly process, whereas existing Indonesian language resources are scarce or non-existant [1]. In addition, although spontaneous recognition is an important component of practical speech recognition, research has only relatively recently shifted in this direction.

Spontaneous and dictated speech also differ in various ways. It has been proposed that spontaneous speech is less ordered both linguistically and acoustically [15], and differs spectraclly from dictated speech [18]. Spontaneous speech contains filled pauses, repairs, hesitations, repetitions, and disfluencies, which are largely absent in dictated speech. The exact differences vary by language, with some languages exhibiting a strong tonal difference between spontaneous and dictated speech [10].

A number of methods have been developed to improve spontaneous speech recognition performance. In the model adaptation approach, a large amount of dictation data is adapted to a small amount of spontaneous data while training the acoustic model, using a method such as Maximum A Priori (MAP) adaptation. In the second approach, acoustic and language models for spontaneous and dictated speech are trained separately, with a method to switch or weigh models on the fly. Model switching relies on a reliable way to differentiate between spontaneous and dictated speech that is independent of the specific acoustic models.

In either case, it is instructive to understand how acoustic differences between spontaneous and dictated speech manifest and how to describe them. With a better understanding of these differences we can train better specialized models for spontaneous or mixed speech, for example utilizing the model switching approach. The goal is to achieve a reasonably performant "front-end" classifier that can pass data to a number of back-end models specialized for the task at hand.

In the following section we describe the relevant literature and the speech corpus used for analysis. We then discuss at length the various phoneme-level differences between spontaneous and read speech in the Indonesian language. Finally we attempt to engineer segment-level acoustic features and utilize the feature set to classify segmented utterances.

II. Related Works

In classifying spontaneous speech, a number of approaches may be taken. Many studies have focused on linguistic features such as grammar, word choice, hesitations, speech rate, pause types, pause structures, intonation, and articulation [3][4][5]. While such features are able to describe the semantic and syntactic characteristics of spontaneous speech, they are limited by the difficulty of labeling the training data. Acoustic characteristics of spontaneous speech, on other hand, are easier to extract, but more difficult to use for classification purposes [18]. Furui has shown that there is a substantial difference in the spectral properties of vowels between read and spontaneous speech [13]. By taking the differences between phoneme vectors of varying styles and their respective centers/averages, the spectral reduction of various phonemes between various

speaking styles can be meaningfully distinguished. A very promising new approach is in extracting the GMM supervectors between multiple utterances [17], though this is limited by the latency of the classification results.

The exact definition of what constitutes "spontaneous" speech may often make or break a particular method. For spontaneous speech that is different from read speech only in terms of preparedness of speech, but controlling for factors such as intonation, pitch, and speaking rate, it is very difficult even for humans to distinguish between the two [9]. An oft-proposed approach to characterizing spontaneous speech is the use of prosodic features that describe the intonation and rhythm of speech [11]. An indirect approach, through the attempt to classify multiple classes, may shed more light into the issues involved [12].

Specifically with regards to Indonesian, recent advances have been made in building large scale systems [2][7]. However problems still persist with regards to spontaneous speech. Hoesen et al., building on previous Indonesian language model adaptation work[6], have developed a reasonably performant Indonesian language speech recognition system, capable of achieving upwards of 80% accuracy for dictated speech but only around 65% for spontaneous speech, even with model adaptation. Methods using general machine learning algorithms have been proposed for handling filled pauses in Indonesian [8].

Detecting spontaneous and read utterances based on unlabelled segments relies on achieving reliable speech segmentation. Many methods have been described in the literature to achieve this, building on the standard Hidden Markov Model (HMM) approach or using newer unsupervised approaches [19]. For this work, the Kaldi speech toolkit [14] is used primarily for speech segment extraction and obtaining the per frame MFCC features of recordings.

III. Speech Corpus

The Perisalah Corpus is a collection of spontaneous and read speech recorded for the development of the Perisalah Speech Recognition System in collaboration with PT. Inti, an Indonesian public telecommunications company. The corpus samples a variety of different age groups and local dialects. The meeting transcription system is being deployed at various government institutions at the national and regional levels. The read/dictation part of the corpus was build to be lexically balanced with respect to written Indonesian.

A. Corpus Overview

After cleaning, the corpus consists of 297 native Indonesian speakers, with on average 275 dictated and 61 spontaneous utterances per speaker, totalling 81240 dictated and 18210 spontaneous utterances, respectively. On average, dictated utterances are 6.54 seconds long and spontaneous utterances 8.33 seconds long for a total of 147.6 and 42.1 hours of dictated and spontaneous speech, respectively.

Demographically, speakers were taken from both genders roughly equally, with 143 male speakers and 154 female speakers. Fig. 1 shows the distribution of genders across dialects, the sampled dialects being Javanese (J), Sundanese (S), Minanga (M), Batak (T), Betawi/Melayu (A), Balinese (B), Sulawesi (W), and Maluku/Papua (P). Roughly 57% of Indonesians live on Java, where the Javanese, Sundanese, and Betawi/Melayu dialects originate. Two age groups were defined; ages 40 years and under, and ages above 40 years, with 243 and 53 speakers, respectively.

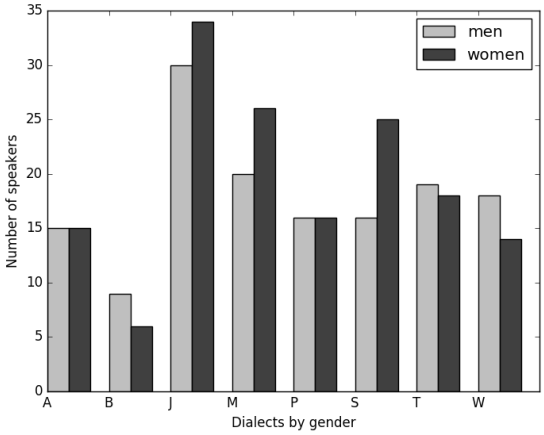

Fig. 1. Number of members of each gender for each dialect

Utterances for prepared texts were taken from newspaper articles and magazines, and hence derived from real-world examples, albeit written ones. A total of ten different sets were compiled, labelled from 'A' to 'J'. Speakers were recorded in a single session, with read speech recorded first and spontaneous speech afterwards. For spontaneous speech, speakers were asked to choose an arbitrary topic in which they were comfortable with and thus asked to speak at length about it.

Fluency of speech varies greatly within the corpus, with some participants experiencing difficulties reading fluently, while others were highly uncomfortable when told to improvise a topic. Utterances for spontaneous speech were segmented by hand; that is, the start and end of sentences is dependent on the studio operator's comprehension of the speech, although supervision was provided. Hence, using the duration of utterances as a metric should be done with caution.

B. Acoustic feature extraction

Speech audio data is represented as audio samples taken at a particular rate. A constant number of possibly overlapping audio samples are processed in a frame to obtain low level spectral, cepstral, and energy features. A segment, comprised of a varying number of such frames, represents a single phoneme. For the purposes of our acoustic analysis, linguistic units other than the phoneme are ignored and we instead consider utterances as sequences of segments. In addition to frame and segment-level acoustic information a number of high level features are obtained at the utterance and speaker levels, namely the specific speaker of an utterance, their gender, dialect, and age.

Segments are force aligned on a prebuilt HMM trained using the Kaldi toolkit. The segment duration, location and phone are obtained for each utterance in the corpus. 171 phones are defined in the corpus, which includes voiced and spoken phones, fillers, silence, and other non-speech phones. The 64 most common phones calculated by frequency of appearance in the corpus are displayed in Table I. These phones account for 98.15% of the segment occurences in the corpus. Spoken phones are further separated as belonging to the beginning, inside, or end of a word, indicated by a 'B', 'I', and 'E' suffix, resepectively. The '@' symbol represents the phonetic 'ə'.

TABLE I
COMMON PHONES

@_B	@_E	@_I	a_B	a_E	a_I	b_B	b_I
c_B	c_I	d_B	d_I	e_B	e_I	f_B	f_I
g_B	g_I	h_B	h_E	h_I	i_B	i_E	i_I
j_B	j_I	k_B	k_E	k_I	l_B	l_E	l_I
m_B	m_E	m_I	n_B	n_E	n_I	ng_E	ng_I
ny_I	o_B	o_E	o_I	p_B	p_E	p_I	r_B
r_E	r_I	s_B	s_E	s_I	sil	t_B	t_E
t_I	u_B	u_E	u_I	w_B	w_I	y_B	y_I

For each frame, the formants and MFCCs are extracted. The first four formants of each frame are calculated by solving its Linear Predictive Coefficients (LPC) an alignment subroutine. The first 13 MFCCs are obtained during the pretraining of the segmentation HMM through the Kaldi toolkit, using the training defaults including Cepstral Mean Subtraction (CMS). The delta and delta-delta features were calculated manually during frame alignment. Apart from averaged frames, features at the segment level include segmental duration and location, and segment label in terms of phoneme. The complete list of features is shown in Table II.

C. Phone analysis

The analysis of data that follows is primarily exploratory, focusing on phone-level differences and with the secondary goal of obtaining feature configurations that maximizes the distance between spontaneous and dictated speech. In this section the data is unnormalized unless where stated to allow for easier interpretation.

A high level view of the data is obtained by first grouping phones into spontaneous and dictated classes and subsequently averaging the segment features in both groups over all utterances and speakers. This view is filtered such that only phones of frequencies above 0.001% in both dictated and spontaneous scenarios are processed further; for a typical speaker, this amounts to six spontaneous occurences of that phone in total. From this view, differences in phone frequency and average duration per phone between spontaneous and dictated speech are noticeable. Table III illustrates this point; the differences in average occurence frequency in percentage points over all common phones between spontaneous and dictated speech are calculated and sorted, with the top and bottom ten results displayed in the upper and lower sections, respectively. This is similarly done for average duration and log energy, with results displayed to the right of the table.

TABLE II
ACOUSTIC FEATURES

Segment-level features

Cepstral (36)	MFCC 12 + Δ12 + ΔΔ12
Log energy (3)	MFCC 1 + 1Δ + 1ΔΔ
Temporal (2)	Duration 1 + Position 1
Spectral (4)	Formants 4 F0 - F4

Utterance-level features

Utterance ID (1)	Utterance number
Label (1)	Spontaneous or dictated 'Z' for spontaneous or one of 'A' - 'J' for dictated

Speaker-level features

Speaker ID (1)	Speaker number
Gender (1)	Female or male
Dialect (1)	Jawa (J), Sunda (S) Minang (M), Batak (T) Betawi/Melayu (A) Bali (B), Sulawesi (W) or Maluku/Papua (P)
Age group (1)	Young (M) or old (L)

TABLE III
MOST DISTANT PHONES IN TERMS OF FREQUENCY, DURATION, AND LOG ENERGY

Phon	Frequency%	Phon	Duration(s)	Phon	Log Energy
e_I	-1.303714	sil	-0.097187	e_B	-5.482111
sil	-0.753674	o_E	-0.066367	a_B	-4.703623
o_I	-0.643983	e_B	-0.038929	@_I	-4.695232
m_I	-0.446265	i_B	-0.031062	e_I	-4.610722
l_I	-0.364059	f_B	-0.030189	a_I	-4.337959
s_I	-0.289272	y_B	-0.030112	o_I	-4.163954
r_I	-0.257801	e_I	-0.028282	l_I	-3.586154
b_I	-0.242627	j_B	-0.027528	i_I	-3.566285
r_E	-0.238013	g_B	-0.027018	r_I	-3.450573
p_B	-0.236948	w_B	-0.026554	u_I	-3.024451
t_I	0.329166	k_E	-0.006259	k_I	0.999814
ny_I	0.368012	l_E	-0.005862	p_I	1.066118
ng_E	0.384281	i_I	-0.004152	o_E	1.138495
@_E	0.432411	@_I	-0.003671	t_B	1.221219
y_B	0.436475	n_E	-0.002305	t_I	1.236945
u_E	0.502933	u_I	-0.001642	p_B	1.387123
@_I	0.508262	h_E	0.007077	k_E	1.421851
@_B	0.525272	ng_E	0.009034	t_E	1.719993
a_E	0.557744	m_E	0.010114	@_B	1.924112
h_E	0.960796	@_B	0.163624	sil	2.854659

Phones higher up on the table exhibit a larger difference between their dictated and spontaneous versions given the particular measure. For example, the phone 'e_I' is overrepresented and 'h_E' underrepresented in the dictated corpus with respect to the spontaneous. The 'sil' phone, in all cases

positioned at the extremes of the table, may be an indicator of uneven quality of data, a result of mislabelled or unlabelled spontaneous speech or differences in the noise floor of audio recordings.

The phones on the upper half of the duration column tend to be situated at the beginning of words while those at the bottom tend not to. The exception to the rule is '@_B', a common filled pause in Indonesian and many other languages, which is voiced longer than its dictated counterpart by the widest margin. The prevalence of word-opening phones at the top and closing phones at the bottom may be indicative of a tendency for Indonesian speakers to rush through the beginnings of words and slow down towards the ends.

Finally, the bottom section of the log energy column is populated mostly by plosives, showing that spontaneous plosives are louder on average than dictated plosives. A possible explanation for this is speakers exerting less conscious control over their speech during spontaneous scenarios. The top two phones, 'e_B' and 'a_B', are in Indonesian often followed by a stressed syllable, which may be less exaggerated in dictation scenarios.

The remaining acoustic features are analyzed to obtain their phone relationships in a similar manner, as seen in Table IV. In analyzing both MFCCs and formants, the log energy components are discarded and the standardized Euclidian distances between spontaneous-dictated phone pairs are aggregated over all segments. The results from formant analysis show only small differences between phones and low variance ($var = 0.201$), hence averaged formant features are not optimal indicators of phone differences between styles. On the other hand, the results of cepstral analysis show large differences but are difficult to interpret correctly.

TABLE IV
MOST DISTANT PHONES IN TERMS OF MFCC AND FORMANTS

Phon	MFCC dist	Phon	Formant dist
k_E	0.962761	l_E	0.435130
sil	1.001989	ng_E	0.467932
t_E	1.003554	@_E	0.472607
g_I	1.201156	l_I	0.474154
s_E	1.277621	h_I	0.486802
g_B	3.254625	u_B	1.924788
u_E	3.319732	g_B	1.973173
y_B	3.424861	e_B	1.979410
u_B	3.507406	i_B	2.063410
@_B	6.020622	@_B	2.625621

For the phonemes for which all three variants are well represented in the data, by assessing the distance between phone variants and their averages, we gain a sense of how important the phoneme's position within the word is towards its acoustic characteristics. The data suggests that position-in-word disproportionately affects the acoustics of plosive and fricative phonemes. This is unsurprising, as in Indonesian fricatives and plosives at word ends are usually unvoiced.

Additional views are generated by grouping phonemes by high level features, namely gender, dialect, and age. The average feature vector for each grouping, phoneme, and style is calculated, the Euclidean distance between each spontaneous-dictated phoneme pair obtained, and the standard deviation of the resulting values is taken. This is interpreted as a measure of the variety in the amount by which spontaneous and dictated speech differs between genders, dialects, and ages. On average, spontaneous-dictation distances vary little between group members, with the exception of distances across dialects, as seen in Table V.

TABLE V
TOP AND BOTTOM STANDARD DEVIATION IN SPONTANEOUS-DICTATED DISTANCES FOR GENDER, DIALECT, AND AGE GROUPINGS

Gender		**Dialect**		**Age**	
Phon	Dist	Phon	Dist	Phon	Age
ny_I	0.000403	t_B	0.047636	t_E	0.012024
e_I	0.002478	h_B	0.048427	ny_I	0.012065
r_E	0.003271	d_I	0.050404	u_E	0.015328
l_E	0.003325	n_I	0.051392	w_B	0.015333
g_I	0.004378	sil	0.051638	h_E	0.018801
m_E	0.189082	g_B	0.212091	g_I	0.167430
e_B	0.213758	c_I	0.225610	k_B	0.171539
c_B	0.225976	o_E	0.242959	o_I	0.172099
y_B	0.228350	f_B	0.305656	u_B	0.196584
i_B	0.269465	e_B	0.394694	l_E	0.216593
Mean	0.008207	Mean	0.125000	Mean	0.822268

Lastly, all segment features are grouped into phonemes, normalized, and mapped to 2-dimensional space by PCA transformation to get a sense of how they relate spatially. Fig. 2 shows a 2-dimensional representation of the complete set of features of all common phones, with their distances represented as dashed lines between pairs. Although distances between points in a pair are discernible, they often overlap with other phone pairs. There is no obvious separation between spontaneous and dictated phones with this feature set.

IV. UTTERANCE CLASSIFICATION

General differences in spontaneous and dictated speech are observable from phoneme analysis. Though useful for gaining insight as to why an existing model performs poorly, it would be practical to separate spontaneous and dictated speech in the absence of linguistic/grammatical information such as phones and (position in) words. In the subsequent sections we shift the focus away from phones and the exploration of overarching differences between spontaneous and dictated speech, towards "anonymous" segments and the task of classifying spontaneous and dictated utterances.

A. Segment Analysis

Intuitively many spontaneous speech traits, such as disfluency and filled pauses, result in less constant speech rates as compared to read speech, with long segments followed by a quick succession of short segments. In order to emulate this,

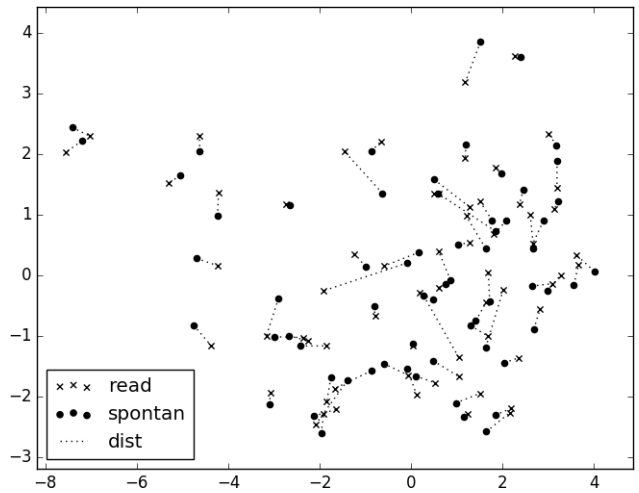

Fig. 2. Distances and positions of dictated-spontaneous phone pairs in 2-D space

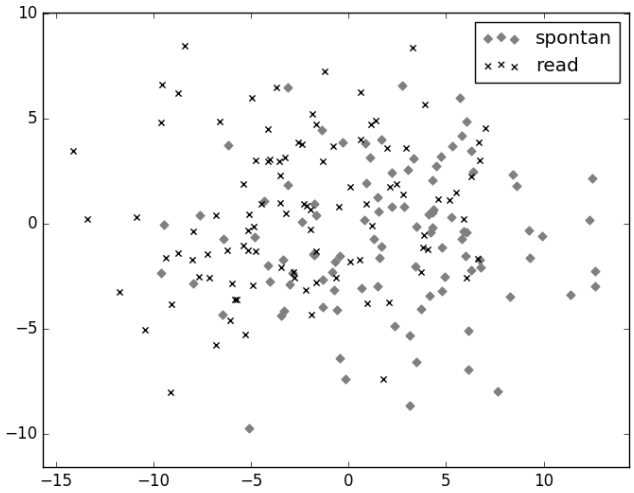

Fig. 3. Utterance samples in 2-D space

a variation of the feature extraction pipeline is implemented whereby for each segment, in addition to the acoustic features, the difference in features against the preceeding segment, henceforth "delta segments", are also extracted. In a manner similar to that of "delta-delta" MFCC features, the delta-delta segment features are also extracted. The purpose of these features is to embed local temporal information at the segment level. The full modified feature set is 264 features long, with both the averages and standard deviations taken for all segment features. Table VI shows the complete list of segment acoustic features.

TABLE VI
MODIFIED SEGMENT-LEVEL FEATURES

	Segment		ΔSegment		ΔΔ Segment	
	Mean	Std	Mean	Std	Mean	Std
Duration	1	1	1	1	1	1
Log energy	3	3	3	3	3	3
Cepstral	36	36	36	36	36	36
Formants	4	4	4	4	4	4

As in the previous section, the formant and MFCC frame features are first averaged over segments. Subsequently, the segment durations are taken, and all three components averaged over the length of the utterance. In addition to the average, the standard deviation is also computed. At the same time, the difference in duration, formant, and MFCC values between the currently evaluated segment and the one immediately preceeding it is calculated and the average and deviation over the utterance computed as the delta segment. This is done once again to obtain the delta-delta segments. In Fig. 3, we map a sample of the utterances extracted this way into two dimensional space by Principle Component Analysis (PCA). Although there is significant overlap in regions, the centers of each are relatively distinct, and analysis in higher dimensions may yield results.

B. Machine Learning and Experimentation

An appropriate method to separate spontaneous speech is required. As delta/delta-delta segments encode temporal information indirectly, it is difficult to utilize a sequence-sensitive method such as a Markov Process. Hence, general approaches such as ensemble methods, Support Vector Machines (SVM), and logistic regression are considered instead.

Ensemble methods employ a large number of weak classifiers built on a particular learning algorithm and combine their predictions in order to improve generalizability and robustness. A random forest is similar to a classification tree, differing in the branching behavior. While in decision/classification tree learning the best predictor (the one that most separates the data) is taken, in random forests a random sample of the predictors is taken instead, and the best split chosen from it [21].

A training pipeline is built utilizing the Python scientific stack, with feature extraction and data analysis methods built on Pandas and cross validation, classification, and evaluation provided by Scikit-Learn [20]. Instead of processing data iteratively, the dataset is processed in parallel utilizing the distributed capabilities of Dask and Pandas.

V. RESULTS AND ANALYSIS

An equal number of positive (spontaneous) and negative (read) utterances are sampled and trained on a number of classifiers. The initial classification results, evaluated by average accuracy, F1 scores, and their respective standard deviations over ten folds, are shown in Table VII. The classifiers considered here are Stochastic Gradient Descent (SGD) classfiers utilizing the Support Vector Machine (SVM) and logistic regression (Log) loss functions, resepectively, and ensemble methods consisting of random forest (R.Forest) and gradient boosting (G.Boost) classification. Ten-fold cross validation on shuffled data is used throughout, with the full feature set utilized to obtain the initial results.

TABLE VII
INITIAL CLASSIFICATION RESULTS

	Accuracy		F1 score	
	Mean	Std	Mean	Std
SVM	0.906142	0.012277	0.906234	0.015051
Log	0.907956	0.007953	0.907332	0.010498
R.Forest	0.928603	0.008885	0.931077	0.007835
G.Boost	0.939792	0.010555	0.938142	0.008998

The nearly identical accuracy and F1 scores for each method indicates that classification accuracy between spontaneous and dictated speech is nearly equal. However, upon closer inspection classification results are consistently worse for spontaneous speech. This is to be expected given the higher variability of spontaneous speech and given that training was conducted on equal amounts of spontaneous and dictated data. On the other hand, this should be somewhat offset by the larger average number of segments per utterance for spontaneous speech in the corpus.

Fig. 4 displays the "hit rate" of the random forest classifier trained previously. This is the squared ratio between the number of hits (correct predictions) and misses (incorrect predictions) given the number of segments in an utterance. Similar results were achieved with the other classifiers. In general, longer utterances produce higher hit rates, indicating classification success. It must be noted that utterance length itself should not be and is not used as a feature, as the length of read utterances are predicated during the design of the prepared text and hence not necessarily indicative of any real world differences between spontaneous and read speech. Still, given the large difference in the number of segments per utterance between read and spontaneous speech in this particular corpus, there is the danger that results will not generalize in real-world cases. Further studies with different datasets must be conducted to investigate the applicability of these findings.

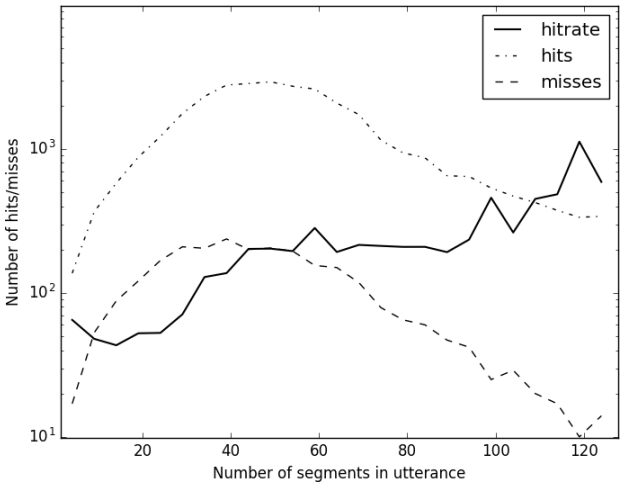

Fig. 4. Number of hits and misses against utterance length

The random forest classifier, possessing the lowest standard deviation of all tested methods, is further tweaked. Parameter tuning is done by first extracting the relative importance of the features utilized previously. The relative importance of a feature measures the fraction of all training examples classified by it. Of the initial training examples, 44.7% of the utterances were classified by the six durational and eighteen log energy features, contributing 24.370%, 9.725%, 7.0291%, and 4.8284% respectively. Training on the reduced feature set leads to a very slight improvement in accuracy (93.27%) and F1 scores (93.23%) and a large reduction in training times. On the reduced feature set, the most relatively important features are the average delta and delta-delta segment duration features at 17.2% and 14.8% respectively, followed by the average delta segment log energy at 7.7% and deviation delta segment duration at 7.0%. Iteratively combining the remaining features and re-training the classifier lead to no further accuracy gains.

VI. CONCLUSION AND FUTURE WORKS

There are distinguishable differences in the acoustic characteristics of Indonesian phonemes between spontaneous and read speech. Measured by phoneme duration, log energy levels, and occurence distribution, there are observable differences in many common vowels and also plosives and fricatives. Exploring methods to better represent these phonemes in training data, both in terms of occurence frequency and acoustic characterisitics, may improve speech recognition performance. In the future, a method to suggest improvements to the existing corpus building process based on these findings would be beneficial.

Preliminary results suggest that delta and delta-delta segment information are promising features in the classification of utterances into spontaneous and dictated Indonesian speech. On a cross-validated training set, F1 scores of 93.23% were achieved on a random forest classifier utilizing a feature set consisting of 24 values; the averages and deviations of segment duration and energy, delta segment duration and energy, and delta-delta segment duration and engery. Classification results above 90% were achieved across all tested classifiers. Further research is needed to determine if this holds true for other datasets and in different acoustic conditions. If robust, this approach may be useful for multi-class classification tasks such as dialect and speaking style classification, or as a front-end to switch between various specialized back-end models.

ACKNOWLEDGEMENTS

This research is partially supported by Riset Unggulan Perguruan Tinggi Kemenristekdikti (University Distinguished Research) and is part of a larger project titled Pengembangan Perangkat Lunak Perisalah Rapat dengan Memanfaatkan Teknologi Pengenal Ucapan, Mesin Translasi, dan Peringkas Otomatis (Development of Meeting Speech Transcriber Using Speech Recognition, Machine Translation, and Automatic Summarizer Technology). We also would like to thank PT INTI, Bandung, Indonesia for the utilization of the speech corpus presented in this research.

REFERENCES

[1] D. Hoesen, C. Satriawan, D.P. Lestari, M.L. Khodra, "Toward Robust Indonesian Speech Recognition with Spontaneous-speech Adapted Acoustic Models," SLTU-2016 5th Workshop on Spoken Language Technologies for under-resourced languages, 2016.

[2] D.P. Lestari, K. Iwano, S. Furui, "A large vocabulary continuous speech recognition system for Indonesian language," Proc. 15th Indonesian Scientific Conference in Japan, 2006.

[3] H.C. Barik, "Cross-linguistic study of temporal characteristics of different types of speech material," Language and Speech 20, pp. 116-126, 1977.

[4] E. Blaauw, "The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech," Speech Communications 14(4), pp. 359-375, 1995.

[5] N. Umeda, K. Wallace, K. Horna, "Usage of words and sentence structures in spontaneous versus text material," Proc. ICSLP vol.1, pp. 759-762, 1992.

[6] D.P. Lestari, A. Irfani, "Acoustic and Language Model Adaptation for Indonesian Spontaneous Speech Recognition," 2nd International Conference on Advanced Informatics: Concepts, Theory, and Application (ICAICTA), 2015.

[7] S. Sakti et al, "Development of Indonesian large vocabulary continuous speech recognition system with A-STAR project," Proc. of the Workshop on TCAST, pp. 19, 2008.

[8] A. Sani, D.P. Lestari, A. Purwarianti, "Filled Pause Detection in Indonesian Spontaneous Speech," Communications in Computer and Information Science vol. 293, pp. 54-64, 2016.

[9] G.P.M. Laan, "The contributions of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," Speech Communication 22:43-65, 1997.

[10] V. Delwo, A. Leemann, M.J. Kolly, "The recognition of read and spontaneous speech in local vernacular: The case of Zurich German," Journal of Phonetics vol. 48, 2015, pp. 13-28.

[11] K. Silverman, E. Blaauw, J. Spitz, J. Pitrelli, "Towards Using Prosody in Speech Recognition/Understanding Systems: Differences Between Read and Spontaneous Speech",

[12] G. Liu, Y. Lei, J.H.L. Hansen, "Dialect Identification: Impact of Differences Between Read and Spontaneous Speech", 18th European Signal Processing Conference (EUISPCO-2010), 2010.

[13] S. Furui, M. Nakamura, T. Ichiba, K. Iwano, "Why is the recognition of spontaneous speech so hard?", 8th Internation Conference on Text, Speech, and Dialog, Karlovy Vary, 2005.

[14] D. Povey, et al., "The Kaldi speech recognition toolkit," IEEE Workshop on ASRU, 2011.

[15] S. Furui, "Recent advances in spontaneous speech recognition and understanding", Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. 1-6, 2003.

[16] M. Benzeguiba, R. DeMori, O. Deroo, S. Dupon, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens, "Automatic Speech Recognition and Speech Variability: a Review", Speech Communication 49:763-786, 2007.

[17] T. Asami, R. Masamura, H. Masataki, S. Sakauchi, "Read and spontaneous speech classification based on variance of GMM supervectors", Interspeech 2014:2375-2379, 2014.

[18] M. Nakamura, I. Koji, S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance", Computer Speech and Language 22(2): 171-184, 2008.

[19] O. Scharenborg, V. Wan, M. Ernestus "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries," Journal of the Acoustical Society of America 127:1084-1095, 2010.

[20] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," The Journal of Machine Learning Research 12 (2011): 2825-2830.

[21] A. Liaw, M. Weiner, "Classification and Regression by randomForest," R News 2(3), 18-22, 2002.