Specific acoustic models for spontaneous and dictated style in Indonesian speech recognition

To cite this article: C B Vista et al 2018 J. Phys.: Conf. Ser. 978 012059

View the article online for updates and enhancements.
Specific acoustic models for spontaneous and dictated style in Indonesian speech recognition

C B Vista¹, C H Satriawan², D P Lestari³ and D H Widyantoro⁴

¹, ², ³, ⁴School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia

¹candrabellavista@gmail.com, ²23515053@std.stei.itb.ac.id, ³dessipuji@stei.itb.ac.id, ⁴dwi@stei.itb.ac.id

Abstract. The performance of an automatic speech recognition system is affected by differences in speech style between the data the model is originally trained upon and incoming speech to be recognized. In this paper, the usage of GMM-HMM acoustic models for specific speech styles is investigated. We develop two systems for the experiments; the first employs a speech style classifier to predict the speech style of incoming speech, either spontaneous or dictated, then decodes this speech using an acoustic model specifically trained for that speech style. The second system uses both acoustic models to recognise incoming speech and decides upon a final result by calculating a confidence score of decoding. Results show that training specific acoustic models for spontaneous and dictated speech styles confers a slight recognition advantage as compared to a baseline model trained on a mixture of spontaneous and dictated training data. In addition, the speech style classifier approach of the first system produced slightly more accurate results than the confidence scoring employed in the second system.

Keywords. Specific model; speech recognition; spontaneous speech; Indonesian language

1. Introduction

A speech recognition system identifies spoken words or sentences from a speech audio stream through a series of computations [1]. Previous GMM-HMM based automatic speech recognition systems for the Indonesian language have produced accurate results [2,3]. However, the accuracy of such systems declines dramatically when recognizing spontaneous speech, such as during conversations, meetings, and interviews, compared to dictated or prepared speech. This could be due to the differences in acoustical and lexical characteristics between spontaneous speech and dictated speech [4]. Spontaneous speech contains repetition, revision, mispronunciation, filled pauses, and hesitation [5]. In addition, spontaneous speech often contains foreign words, typically English, which complicates recognition by incurring errors expressed as specified terms, affixation, and new phonemes influenced by the speaker’s dialect, a common occurrence in Indonesian conversation [3].

Compatibility between the trained acoustic model and language model with input speech greatly affects the accuracy of speech recognition. Several previous studies have attempted to build models that are more robust towards the differing characteristics of input speech through model adaptation [6]. To investigate the effect of model compatibility with speech input, some experiments to build specific models for spontaneous speech and dictated style were conducted.

2. Related Works
In previous research, models trained for specific speaker age groups led to a rise in recognition accuracy for Italian language spontaneous and dictated speech [7]. The usage of separate age-group-specific models led to a 0.4% and 0.7% decrease in word error rate (WER) when tested against adult and child speech data, respectively, over a baseline system utilizing a single model.

In [8], a Czech language automatic speech recognition and subtitling system was built using separate speech style models. This study used spontaneous speech training data from 64 male and 36 female speakers. The training of separate models for each gender led to a decline in the WER rate by 2%.

Based on previous research, the training of separate acoustic models for specific groupings promises improved speech recognition performance. In this study, we build specific acoustic models for two speech styles, namely spontaneous and dictated speech, then test this approach by decoding spontaneous and dictated speech with the corresponding model. The results of this approach are then compared to the baseline system consisting of a single acoustic model for both speech styles.

The differences in spontaneous and dictated speech vary by language. A previous study [9] analyzed the acoustic characteristics of Indonesian language phonemes in spontaneous speech by phoneme duration, log energy, and frequency of occurrence. From a total of 33 analyzed phonemes in the Indonesian language, the occurrence rate of the @ phoneme, or ‘ə’ in the International Phonetic Alphabet (IPA), is significantly higher in spontaneous speech as it is commonly used as a filled pause in the Indonesian language. The average duration of phonemes towards the beginning of words is higher than towards the end of words in spontaneous speech as compared to dictated speech, indicating that during spontaneous speech speakers tend to slow down towards the end of words. The phonemes ‘a’ and ‘e’ are more frequently followed by certain strain in spontaneous speech thus phoneme’s log energy of a and e (e in the word of ‘enak’) is higher compared to another phoneme.

3. Speech Corpus Description
For acoustic model training, a speech corpus containing both spontaneous and dictated speech were used. The speech corpora has been previously used in [6]. 300 speakers were recorded for the corpus, with each speaker uttering 300 dictated sentences from a prepared text and a variable number of spontaneous sentences from topical prompts. For the spontaneous portion, utterances were derived from longer statements by manual segmentation. The corpus was recorded in a studio acoustical environment at 16 kHz sampling rate and 16 bit resolution. In addition to speaker identity, each utterance further contained metadata for speaker gender, locale (assumed to influence dialect), and age group. Table 1 summarizes the statistics of the speech corpus used in this work.

| Type       | Spontaneous | Dictated |
|------------|-------------|----------|
| Speakers   | 299         | 300      |
| Sentences  | 18806       | 83310    |
| Duration   | 74 hours    | 150 hours|

4. Experiment Set Up
The acoustic models for the experiments are GMM-HMM-based. Feature extraction, training, and decoding was handled using the Kaldi toolkit [10]. The language model and lexicon from [6] throughout training and testing. The language model utilizes the standard 3-gram form. The text corpus used to build the language model consists of 613.054 sentences and spontaneous and dictated speech transcriptions. The lexicon was built from words contained in the text corpus, taking into account the frequency of words. Each word listed in the lexicon may have one or more alternative pronunciations due to differences in dialect contained in the results of spontaneous and dictated speech transcription.

Two systems were designed and built for the experiment. The systems share a common speech recognition system consisting of three separate acoustic models, namely a spontaneous-specific
acoustic model, a dictation-specific acoustic model, and a baseline acoustic model that uses a mixture of both spontaneous and dictated training data. In addition to the common speech recognition system, the first system utilizes a speech style random forest classifier, the details of which can be found in [9]. Based on the prediction results of the speech style classifier, each input testing utterance is decoded using the appropriate acoustic model. In contrast, the second system decodes each input testing utterance with each model in the common speech recognition system. It subsequently calculates a confidence score to predict which decoding process best predicted the input utterance. If the confidence score for decoding using the speech style specific models is higher than the score for decoding using the integrated model, than the prediction of the speech style specific models is used, and vice versa. Figure 1 and figure 2 shows the two architecture systems.

![Figure 1. Architecture system of automatic speech recognition system using spoken style classifier](image1)

![Figure 2. Architecture system of automatic speech recognition system using confidence score](image2)

During experimentation, an initial set of results are obtained in which each separate style-specific model is tested against speech data of the corresponding type. By comparing the recognition results of the speech style-specific models against the baseline (mixture of spontaneous and dictated) model, we ascertain whether such separate modeling is beneficial. These results also determine the theoretical upper bound of style-specific recognition, which occur when the predictions of the preceeding speech style classifier or of the subsequent confidence score calculation, as in the systems above, are perfectly accurate. This is under the assumption that all spontaneous speech is more accurately recognized using a spontaneous-specific model, and vice versa for dictated speech. Table 2 summarises the relevant combinations. Each separate style-specific model is tested against the appropriate speech style, while the integrated model is tested against all data.

| Acoustic Models       | Test Data                        |
|-----------------------|---------------------------------|
| Spontaneous (S)       | Speech Corpus                    |
| Dictated (D)          | Text Corpus                      |
| Spontaneous + Dictated (S+D) |                      |

Table 2. Experimental Design of Specific Acoustic Model
5. Result

Initially, the speech recognition system is tested without the preceding speech style classifier nor the subsequent confidence scoring as in the systems described previously. In this initial experiment and in all subsequent ones, the speech data from the 300 speakers in the dictated portion of the corpus and the 299 speakers of the spontaneous portion of the corpus are divided into ten subsets of data for 10-fold cross validation. Tests were performed ten times for each set of experiments, and the average value of the Word Error Rate (WER) as a percentage was calculated from these results. The experiment results are presented in Table 3.

| Test Data       | Model | Monophone | Triphone |
|-----------------|-------|-----------|----------|
| Spontaneous (S) | S     | 49.5      | 34.4     |
|                 | S+D   | 51.2      | 34.9     |
| Dictated (D)    | D     | 15.0      | 8.2      |
|                 | S+D   | 16.2      | 8.6      |

When spontaneous speech data is decoded using the spontaneous monophone model, a WER of 49.5% is achieved, a 1.7% improvement over the baseline monophone model whereby spontaneous speech is decoded using the integrated S+D model. The same scenario conducted using the triphone model instead of the monophone results in a WER of 34.4%, a slight 0.5% improvement over the corresponding S+D model.

In the case of dictated speech, on average recognition error rates were lower than for spontaneous speech. Decoding dictated speech using the dictation-specific triphone and monophone model produced WERs of 15.0% and 8.2%, respectively. Decoding dictated speech using the mixed S+D triphone and monophone models produced WERs of 16.2% and 8.6%, respectively, a 1.2% and 0.4% decline against the dictation-specific triphone and monophone models, respectively.

By comparing the recognition results of style-specific models against their mixed model counterparts, we may conclude that on average employing style-specific models leads to an improvement in recognition results, although this improvement is usually slight. In addition, it must be considered that these results assume a perfect method to identify the speech style or determine the “better” result, and inaccuracies in the identification of speech style may offset any improvements in recognition gained by employing style-specific acoustic models. In the proceeding experiments we test the speech style classifier and the confidence scoring algorithm separately, and subsequently, in tandem with the speech recognition system.

The speech style classifier in [8] employs a random forest classifier to categorize the speech input as spontaneous or dictated. Before the decoding process, testing data will be recognized in advance by the spoken style classifier, which is labeled as either spontaneous or dictated speech. The test data that has been labeled as spontaneous or dictated is then decoded using the appropriate acoustic model. Table 4 shows the classification results of the dictated test data and spontaneous test data. The classification results of the dictated test data shows that 14.7% of the test utterances were misclassified as spontaneous utterances. In comparison, 12.3% of spontaneous test utterances were misclassified as being dictated utterances.

| Test Data  | Total Test Data | Misclassified | % Misclassified |
|------------|-----------------|---------------|-----------------|
|            |                 |               |                 |

Table 3. Result of Experiments on Specific Acoustic Model

Table 4. Classification Result
Evaluation of the confidence scoring method was conducted by decoding utterances of a certain style with both its style-specific model and the baseline S+D model, calculating confidence scores, and counting the number of utterances whereby the confidence score of decoding with the baseline model was higher than with the style-specific model. The results of the confidence score evaluation are displayed in Table 5 below. Classification using confidence scores results in a 37.7% and 52.2% misclassification rate for dictated and spontaneous speech, respectively. This “misclassification rate” signifies the number of utterances whereby the confidence score of utilizing the style-specific model for decoding was lower than utilizing the baseline S+D model.

Table 5. Classification Result

| Test Data         | Total Test Data | Misclassified | % Misclassified |
|-------------------|-----------------|---------------|-----------------|
| Dictated (D)       | 83310           | 2961          | 37.7            |
| Spontaneous (S)   | 18806           | 986           | 52.2            |

Finally, we experiment on the full systems as described in the previous section; the first system prepended by the speech style classifier and the second system appended by the confidence scoring algorithm. By comparing the initial baseline model experiments with the system evaluation results using speech style classification and confidence scores, we obtain the results shown in Table 6 below.

Table 6. Classification Result

| Acoustic Models       | Spontaneous Test Data | Dictated Test Data | Average |
|-----------------------|-----------------------|--------------------|---------|
| Mixture of spontaneous and dictated (S+D) | 34.9 | 8.6 | 21.8 |
| Spoken Style Classifier | 34.0 | 8.0 | 21.0 |
| Confidence Score      | 34.6 | 8.3 | 21.5 |

As compared to the mixed S+D model, the utilization of style-specific models in the two systems are shown to improve speech recognition results. The first system, employing the speech style classifier, produced WERs of 34.0% and 8.0%, a 0.9% and 0.6% improvement over the baseline model for spontaneous and dictated test data, respectively. The second system, utilizing confidence scoring, produced WER values of 34.6% and 8.3%, an improvement of 0.3% in both cases over the baseline S+D model for spontaneous and dictated speech, respectively.

These results show that the usage of style-specific models leads to increased recognition performance. On average, a 0.8% and 0.3% improvement in WER over recognition using a sole baseline model trained on both styles of speech was achieved with the usage of a speech style classifier and confidence scoring, respectively, in conjunction with separate style-specific models.

It can be seen that the speech style classifier appears to achieve recognition rates exceeding the theoretical upper bound determined through our initial results; the speech style classifier achieves 34.0% and 8.0% WER for spontaneous and dictated speech, respectively, whereas our initial results suggest 34.4% and 8.2% WER as the upper bound for spontaneous and dictated speech, respectively. This perceived discrepancy warrants further research, but can be explained by the fact that some spontaneous utterances were better recognized utilizing the dictation-specific model, and vice versa. With this in mind, we may come to the conclusion that parts of the training data were “mislabeled”, in the sense that some “spontaneous” utterances were spoken in a manner more strongly resembling...
dictated speech, and vice versa. Further research is needed to determine whether this occurs at the utterance level or at the speaker or other group level.

6. **Conclusion and Further Work**

The results show that the usage of style-specific acoustic models for spontaneous and dictated speech improves speech recognition performance, as compared to a baseline model trained on a combination of spontaneous and dictated speech data. Through initial experimentation, it was shown that decoding spontaneous and dictated speech using a corresponding style-specific triphone model led to a decline in Word Error Rate (WER) of 0.6% and 0.2%, respectively.

The usage of a speech style classifier in conjunction with these separate models resulted in a 0.9% and 0.6% decline in WERs for spontaneous and dictated speech, respectively, over the baseline system. The usage of confidence scoring following the decoding of incoming speech through the separate models led to a 0.8% and 0.3% decline in WERs for spontaneous and dictated speech, respectively, over the baseline system.

Further research is needed to determine whether the use of style-specific models improves recognition for other style groupings aside from binary spontaneous/dictated groups. In addition, the use of hybrid or ensemble approaches to classifying speech styles, for instance employing both speech style classification and confidence scoring, improves recognition results.

**References**

[1] Jurafsky D and Martin J H 2008 Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition. Upper Saddle River, NJ: Prentice-Hall.

[2] Sakti S, Kelana E, Riza H, Sakai S, Markov K, and Nakamura S 2008 Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project. *Proceedings of the Workshop on TCAST* pp 19-24.

[3] Lestari D and Furui S 2010 Adaptation to Pronunciation Variations in Indonesian Spoken Query-Based Information Retrieval. *IEICE Transactions on Information and System* 93-D(9):2388-2396.

[4] Furui S, Nakamura M, Ichiba T and Iwano K 2005 Why is the Recognition of Spontaneous Speech so Hard. 8th International Conference Text, Speech, and Dialogue, (Karlovy Vary, Czech Republic) pp 9-22.

[5] Nakamura M, Iwano K, and Furui S 2008 Differences between Acoustic Characteristics of Spontaneous and Read Speech and Their Effects on Speech Recognition Performance. *Computer Speech and Language* pp 171-184.

[6] Hoesen D, Lestari, D P and Khodra M 2016 Adaptation of Acoustic Model for Indonesian Using Varying Ratios of Spontaneous Speech Data. 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA), Bali, Indonesia.

[7] Gerosa M 2009 Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech. 2193-96.

[8] Vanek J and Psutka J 2010 Gender-Dependent Acoustic Models Fusion Developed for Automatic Subtitling of Parliament Mettings Broadcasted by the Czech TV. *International Conference on Text, Speech, and Dialogue, Springer Berlin Heidelberg* pp 431-38.

[9] Satriawan C and Lestari D 2016 Analyzing Differences in Indonesian Spontaneous and Dictation Speech. Bali, Indonesia.: 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment

[10] Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Vesely K, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G 2011 The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.*
Acknowledgement
This research was partially supported by the Master’s Program toward Doctoral Degree for Excellent Graduate (Program Pendidikan Magister Menuju Doktor untuk Sarjana Unggul/PMDSU) from Kemenristekdikti Indonesia within research entitled “Indonesian Automatic Speech Recognition System”.
