論文 / 著書情報
Article / Book Information

論題(和文)	
Title	Automatic recognition of Indonesian declarative questions and statements using polynomial coefficients of the pitch contours
著者(和文)	篠田 浩一, 古井 貞熙
Author	Nazrul Effendy, Koichi Shinoda, Sadaoki Furui, Somchai Jitapunkul
出典(和文)	, Vol. , No. 30, pp. 249-256
Journal/Book name	The Acoustical Society of Japan, Accoust. Sci. & Tech., Vol. , No. 30, pp. 249-256
発行日 / Issue date	2009, 4

Acoust. Sci. & Tech. **30**, 4 (2009)

PAPER

Automatic recognition of Indonesian declarative questions and statements using polynomial coefficients of the pitch contours

Nazrul Effendy[1,2,*], Koichi Shinoda[3,†], Sadaoki Furui[3,‡] and Somchai Jitapunkul[1,§]

[1]*Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University*
[2]*Department of Engineering Physics, Faculty of Engineering, Gadjahmada University*
[3]*Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology*

(*Received 29 July 2008, Accepted for publication 7 January 2009*)

Abstract: We propose an automatic utterance type recognizer that distinguishes declarative questions from statements in Indonesian speech. Since utterances in these two types have the same words with the same order and differ only in their intonations, their classification requires not only a speech recognizer, but also an intonation recognizer. In this paper, the most important utterance part for distinguishing those two types is first identified by perceptual experiments. Then, an utterance type recognizer using that part is proposed, where polynomial expansion is used as a feature extractor and a neural network is used as a classifier. We evaluated our method using Indonesian speech database including 29 pairs of sentences of those two types, each of which uttered by 35 speakers. It was proved that *final word* and *final-two-syllables* are equally effective for the discrimination of each utterance. The proposed recognizer achieved the best accuracy of 89.1% when the order of polynomial expansion was three and the neural network was a linear perceptron.

Keywords: Utterance type, Prosody, Recognizer, Neural networks

PACS number: 43.66.Hg, 43.72.Bs, 43.72.Ne [doi:10.1250/ast.30.249]

1. INTRODUCTION

Usually utterances are grouped into four types: questions, statements, exclamations, and commands. Utterance types are also called utterance classes, dialogue moves, dialogue acts, and speech acts [1]. Utterance type information is important for speech understanding. Spoken dialogue systems often use it to identify users intention [2]. Automatic speech recognition systems use it to decrease the word error rate of the system [3–7]. Speech translation systems use it to resolve ambiguities in translating utterances [8–10].

The question type is further divided into two types: open-question and yes-no-question [11]. An open question involves question words such as *where*, *how*, *why*, etc. A yes-no question allows only two possible responses, positive (*yes*) or negative (*no*). In Indonesian language, a yes-no question is generated in three ways; (1) by using the question indicator 'apa' with or without the interrogative suffix '-kah', (2) by using the interrogative '-kah', and (3) by intonation [12]. The last type (3) is called a declarative question (DQ). Since a DQ has the same words in the same order as a statement, it is impossible to distinguish them only from their transcriptions. Automatic discrimination between a DQ and a statement in Indonesian utterances needs not only a speech recognizer, but also an intonation recognizer, but to the best of our knowledge, no studies have ever addressed this problem.

The intonation of DQs and statements in Indonesian language are subject to change from various reasons [13–16]. A speaker can utter a word of Indonesian with different stressed syllables without changing its meaning [12,17–19]. Even native listeners are often confused in recognizing the two utterance types in some Indonesian dialects, which are especially the dialects that differ from their own dialects. These phenomena make the utterance type identification using the intonation information a difficult task. We believe that a statistical approach based on Indonesian speech data is more robust against these variations than a rule-based approach using some rules built using prior knowledge about the language.

*e-mail: nazrul@gadjahmada.edu
†e-mail: shinoda@cs.titech.ac.jp
‡e-mail: furui@cs.titech.ac.jp
§e-mail: somchai.j@chula.ac.th

Several methods have been investigated to recognize the utterance types in other languages. Those methods include n-gram language models, hidden Markov models, nave Bayes classifiers, Bayesian networks, multilayer perceptrons, decision trees, transformation-based learning, and memory-based learning [3,5,9,20–25]. Most of them, however, utilized language model information, which cannot be used in our problem since the two types, DQs and statements, have the same transcription. A few studies utilized pitch contour representation for utterance type classification. For example, Ishi *et al.* used perceptually-related *F*0 parameters to automatically classify phrase final tones of Japanese [25] and proved its effectiveness. Since it heavily relied on the syllable structure of Japanese, it cannot be used for our purpose. Effendy *et al.* (2004) analyzed the utterance types in Indonesian speech. They used Fujisaki model to represent DQs and statements in Indonesian speech [13]. However, the algorithm to estimate the parameters of Fujisaki model was too complicated to be implemented in an automatic utterance type recognizer and may not be applicable to speaker-independent recognition.

In this paper, we propose a speaker-independent utterance type recognizer to distinguish DQs from statements in Indonesian speech. We first conduct perceptual experiments to reveal the most prominent of the utterance parts that distinguish DQs and statements. Four types of the utterance parts: *final word*, *final syllable*, *final-two-syllables*, and *final-but-one-syllable* (the syllable before *final syllable*) are investigated in the experiments. Then, we use the pitch contour information, which is the major correlate of intonation [26–28] and apply the polynomial expansion [29–31] to extract features from the pitch contour. Finally, a neural network classifier is used to distinguish between DQs and statements in Indonesian speech. This method is expected to be robust against variation in intonation and speaker characteristics.

The next section describes the database used in this research. Section 3 explains the perceptual experiment, Section 4 explains our recognizer, Section 5 reports the results of our evaluation, and Section 6 concludes this paper.

2. SPEECH DATA

We collected speech data of 29 sentence pairs, each of which consists of a statement and its corresponding DQ. The two sentences in each pair have the same words in the same order and differ only in intonation. The sentences are selected from the daily life conversation among Indonesian speakers and listed in Table 1 [13]. We recorded speech data of 35 Indonesian native speakers. They consist of 11 female and 24 male speakers with ages ranging from 23 to 50. The recordings of the speech data were carried out in an office environment. In this recording, we asked each subject to utter the 29 pairs of sentences as naturally as possible. Unlike what Yuan *et al.* did [32], the number of words in each sentence differed from each other. After recording, two subjects verified the speech data subjectively three times and removed the utterances with wrong intonation. In the experiment, 112 utterances were removed. Therefore, there were 1,918 utterances that were available to investigate the utterance type recognizer.

Then, the speech data were divided into four sets, keeping the balance in gender and data amount as illustrated in Table 2. Set I and Set IV contain speech data from 12 male speakers and 5 female speakers. Set II and Set III contain speech data from another group of speakers; 12 male speakers and 6 female speakers. Set I and Set II contain speech data consisting of 14 pairs of sentences, while Set III and Set IV contain speech data consisting of 15 pairs of sentences from another group. Consequently, there are 1866 utterances consisting of 221 statements and 221 DQs in set I, 234 statements and 234 DQs in set II, 241 statements and 241 DQs in set III, and 237 statements and 237 DQs in set IV. In the balance process between statements and DQs in the same set, we removed 12 statements in set I, 10 statements in set II, 22 statements in set III, and 8 statements in set IV.

3. PERCEPTUAL EXPERIMENT

Speech data uttered by four male and four female speakers, which are chosen from the speech data that are explained in Section 2, are used in the perceptual experiment. The experiment is aimed at finding the most prominent of the utterance parts that distinguish DQs and statements in Indonesian speech. We pre-process the speech data before using them in the experiment. In the pre-processing, we manually segment the data to get specific parts of utterances: *final word*, *final syllable*, *final-two-syllables*, and *final-but-one-syllable*. In the segmentation process, we listen to the wave file of each sentence to look for the boundary of the specific parts. Then, we manually cut the wave file to extract the specific parts. From the segmentation of female speech data, we have 230 *final words*, 231 *final syllables*, 153 *final-two-syllables*, and 234 *final-but-one-syllables*. From the segmentation of male speech data, we have 231 *final words*, 228 *final syllables*, 148 *final-two-syllables*, and 227 *final-but-one-syllables*. The data are described in Table 3. The number of each type of the utterance parts is not the same because the speech data consist of final words that consist of different numbers of *final syllable*, *final-two-syllables*, and *final-but-one-syllable*, as a natural characteristic of Indonesian speech.

After preparation and segmentation of the speech data, we asked five male subjects, which are different with the speakers of the speech data, to listen to the speech data and to try to recognize the type of the utterances.

Table 1 Twenty nine sentences selected from the daily life conversation among Indonesian speakers.

No	Sentence in Indonesian	Translation in English
1	Tikar itu baru saja dicuci	That mat was just cleaned
2	Tembok itu dikotori oleh Iwan	That wall was dirtied by Iwan
3	Jagoan itu menendang tiga orang penjahat	That hero kicks three criminals
4	polisi menangkap penjahat pagi tadi	A policeman catches a criminal this morning
5	Sepeda Iwan masih di bengkel	Iwan's bicycle is still in a service center
6	Dia sudah pergi tadi pagi	He has gone this morning
7	Pensil ini sudah tidak runcing	This pencil has not been sharp
8	Gelas ini mudah pecah	This glass is fragile
9	Dia suka menolong orang lain	He likes to help other peoples
10	Lisa sedang menyanyi dan menari	Lisa is singing and dancing
11	Dia sudah makan	He already eat
12	Albert lupa pada dirinya sendiri	Albert forget on himself
13	Kunci pintu itu dibobol maling	The key of that door is stolen
14	Kucing telah menangkap seekor tikus	A cat has caught a mouse
15	Komputer itu terjangkit virus	That computer is infected by virus
16	Ibu sedang belanja ke pasar	Mother is shopping to a market
17	Jendela itu tidak bisa dibuka	That window cannot be open
18	Headphone itu sangat bagus	That headphone is very good
19	Dokter sedang memeriksa pasien	The doctor is diagnosing a patient
20	Kursi ini baru dicat	This chair is just painted
21	Atap rumahnya sudah bocor	The roof of his house has been broken
22	Lantai itu sudah kamu bersihkan	The floor has been cleaned by you
23	Adik sedang bermain	The younger brother is playing
24	Andi berlari ke arah mobil	Andi run to a car
25	Kucing dan anjing sedang berkelahi	A cat and a dog are fighting
26	Televisimu telah dibeli oleh Umar	Your television has been bought by Umar
27	Samsul belum selesai membaca buku itu	Samsul has not finished reading that book yet
28	Bapak sedang mengajar Matematika	Father is teaching mathematics
29	Air di tangki sudah penuh	Water in the tank has been full

Table 2 The Indonesian speech database of statements and declarative questions.

Sentences	Speakers	
	First group of speakers - 12 male speakers - 5 female speakers	Second group of speakers -12 male speakers - 6 female speakers
14 pairs of sentences in the first group	Set I Statements: 221 Declarative questions: 221	Set II Statements: 234 Declarative questions: 234
15 pairs of sentences in the second group	Set IV Statements: 237 Declarative questions: 237	Set III Statements: 241 Declarative questions: 241

Table 4 shows the correctness of the utterance type identification. The experimental results indicated that the subjects recognized the two utterance types with various recognition rates depending on what the utterance part they listen to. The average recognition rate by listening only to *final-but-one-syllable* is the lowest: 72.5%. The average recognition rates of the subjects by listening to the other parts of the utterances: *final word*, *final syllable*, and *final-two-syllables* are more than 90.0%, which are much higher than the average recognition rate by listening to *final-but-one-syllable*. Three listeners identified the sentence types using final-two-syllables with higher recognition rate than using *final word*. However its average recognition rate, 96.8%, is lower than the average recognition rate using final word, 97.1%. Based on the experimental results, we conclude that the *final word* and the *final-two-syllables* are both important utterance parts for the discrimination. In our preliminary experiments using automatic utterance recognizer under the assumption the syllable boundaries are given, these two parts had the similar performance (the *final-two-syllables* 0.1% better than *final word* in recognition accuracy). We prefer to use *final word* in this study,

Table 3 Speech data for the perceptual experiment.

Speech Data	# Utterances	
Female		
final word	230	
final syllable	231	
final-two-syllables	153	
final-but-one-syllable	234	
Total Female		848
Male		
final word	231	
final syllable	228	
final-two-syllables	148	
final-but-one-syllable	227	
Total Male		834
TOTAL		1682

Table 4 The correctness of the utterance type recognition by the listeners in the perceptual experiment. FW: *final word*, FS: *final syllable*, FTS: *final-two-syllables*, FBOS: *final-but-one-syllable*.

Listeners	Correctness (%)			
	FW	FS	FTS	FBOS
S1	97.4	95.4	96.4	76.8
S2	95.5	95.0	98.4	70.0
S3	98.7	97.8	100.0	75.7
S4	98.5	95.9	89.4	68.8
S5	99.2	73.9	100.0	71.2
Average	97.9	91.6	96.8	72.5

however, mainly because the word boundary is more accurately obtained than the syllable boundary by our segmentation module.

4. AUTOMATIC UTTERANCE TYPE RECOGNIZER

We propose an automatic utterance type recognizer to distinguish statements and DQs in Indonesian speech. This recognizer is speaker- and gender-independent, and consists of an automatic utterance segmentation module, an $F0$ extractor, a normalizer, a feature extractor, and a classifier as illustrated in Fig. 1. In this study, we assume that the correct transcription is given, since there are no large databases available to train accurate acoustic models in Indonesian language at present. Each of the subsystem of the recognizer will be described further in the next subsection.

4.1. Automatic Utterance Segmentation

To get the final word boundary, we design an automatic utterance segmentation module using HTK [33]. We assume that the transcription of each utterance is known, and perform alignment between each utterance and its transcription using Viterbi algorithm. For this procedure,

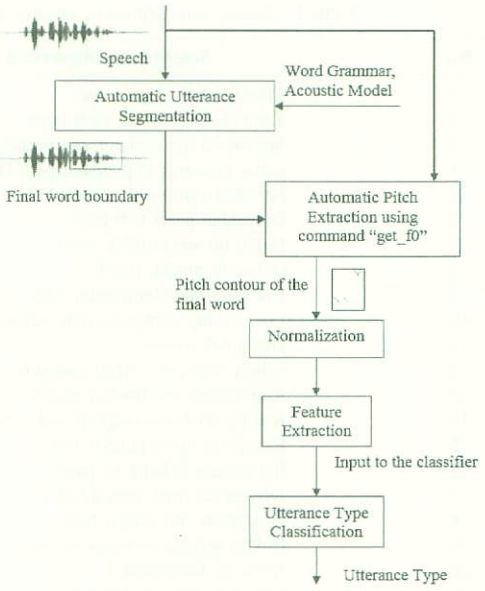

Fig. 1 Block diagram of the proposed automatic utterance type recognizer.

we use a standard feature set in speech recognition and make an acoustic model using a relatively small amount of training data, which is different from the database described in Section 2. The detailed conditions will be explained in Section 5.

4.2. *F0* Extractor

We use the 'get_f0' program from the Entropic Waves software package [34,35] to extract $F0$ data from the final word in each utterance. The get_f0 implements a fundamental frequency estimation algorithm using a normalized cross correlation function and a dynamic programming function. In the experiments described in the next section, we used the default values of the parameters of the 'get_f0', i.e., Gaussian window with the length of 40 ms, and the shift time of 10 ms. The $F0$ data are converted into logarithmic scale and passed through a normalizer.

4.3. Normalizer

Even for the utterances of the same sentence with the same utterance type, the pitch contour of their final word may be different from speaker to speaker. This difference increases the variation in the pitch contour. To achieve robustness against this variation among speakers, the $\log F0$ values are normalized both in the frequency domains and in the time domains. Figure 2 shows the histogram of the duration per syllable of Indonesian statements and DQs. The figure indicated that the time

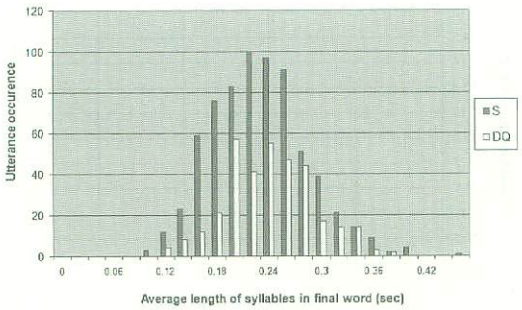

Fig. 2 The histogram of the average length of syllables of the *final word* of Indonesian statements (S) and declarative questions (DQ).

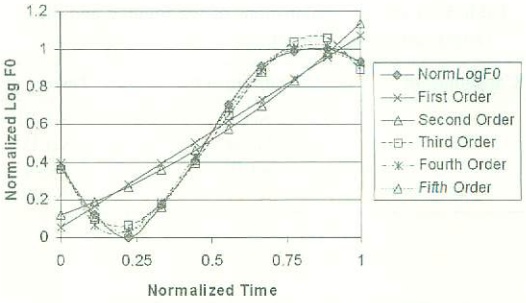

Fig. 3 Representation of typical $F0$ data using various orders of polynomial expansion.

duration is not important in distinguishing Indonesian statements and DQs.

Let p_i $(i = 1, 2, \ldots, L)$ and t_i $(i = 1, 2, \ldots, L)$ be the sequences of the $\log F0$ values and the time of the final word with length L, respectively. Then, the normalized pitch frequency of p_i, $\tilde{p}_i$ $(i = 1, 2, \ldots, L)$ is calculated as:

$$\tilde{p}_i = \frac{p_i - p_{\min}}{p_{\max} - p_{\min}}, \tag{1}$$

where $p_{\max}$ and $p_{\min}$ are the maximum and the minimum $\log F0$ values of the final word. The normalized time of t_i, $\tilde{t}_i$ $(i = 1, 2, \ldots, L)$ is calculated as:

$$\tilde{t}_i = \frac{t_i - t_1}{t_L - t_1}. \tag{2}$$

We normalized the time duration of the utterance part because our experiment implied that the information is not important to distinguish between statements and DQs. Figure 2 shows the histogram of the average length of syllables in the final word of Indonesian statements and DQs. The normalized $\log F0$ values are the input of the feature extractor.

4.4. Feature Extractor

We extract features of the $F0$ contour using the polynomial expansion method [29], where the pitch contour is approximated as a polynomial line in two-dimensional plane of the normalized $\log F0$ and time. The coefficients ci of the polynomial expansion are extracted using the least mean square (LMS) algorithm. Since our method is based on statistical approach, it is expected to be robust against variation in intonation.

Let N be the order of the polynomial expansion. Then the approximated contour for the normalized $\log F0$, $\hat{\tilde{p}}$, is expressed as

$$\hat{\tilde{p}} = \sum_{i=0}^{N} c_i \tilde{t}_i. \tag{3}$$

We remove the coefficient for $i = 0$ to achieve robustness against the difference in the pitch level between male and female speakers.

Figure 3 shows the representation of typical $F0$ data using various orders of polynomial expansion. The higher the order of the polynomial expansion, the smaller the error between the estimated points and the original $F0$ data. As will be explained in the next section, however, the best performance of the automatic utterance type recognizer may not be achieved with the highest order of polynomial expansion because of the over-training problem.

4.5. Classifier

We use a linear perceptron [36] as a classifier in the automatic utterance type recognizer. The number of nodes in the input layer is equal to the number of features extracted from the $F0$ contour. We use one node in the output layer, which is trained to output zero for the statement and one for the DQ. For comparison, we use also a multilayer perceptron with various numbers of hidden layers, as illustrated in Fig. 4.

A statement-DQ threshold is utilized at the end of the output node to classify the utterance type. The threshold is controlled in each experiment such that the error rates for both classes are equal.

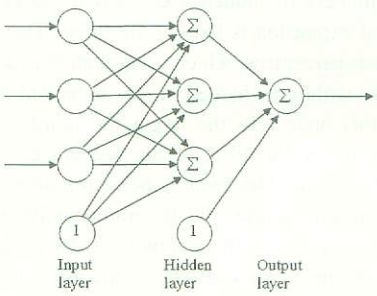

Fig. 4 A multilayer perceptron.

Table 5 The four combinations of the training and the testing sets for the evaluation.

Combination	Training set	Testing set
1	Set I	Set III
2	Set II	Set IV
3	Set III	Set I
4	Set IV	Set II

5. PERFORMANCE EVALUATION

5.1. Experimental Conditions

For the automatic utterance segmentation described in Subsection 4.1, we constructed an acoustic model in the following procedure. First, we recorded 3,840 utterances from 11 male and 9 female speakers, in which each speaker read texts from Indonesian newspapers and Indonesian linguistics books. The texts were chosen in order to contain all phonemes that appear in Indonesian speech. Then, for each 10 ms frame, we extracted features of the power and 12 mel-frequency cepstral coefficients (MFCC), and their first and second order derivatives. The total dimension of the feature vector was 39. Finally, using the training data, we trained monophone hidden Markov models (HMMs) with five states for each phone and 16 Gaussian mixtures for each state.

Using the data sets of speech data as listed in Table 2, we designed four combinations of training and testing sets. The combinations are illustrated in Table 5. The classifier is trained for 100,000 epochs using a back-propagation algorithm. The performance of the automatic utterance type recognizer was evaluated by the averaged equal error rate (EER), which is the average of the EERs of the four combinations in Table 5.

5.2. Experimental Results and Discussions

First, we designed the structure of the neural network classifier. We fixed the number of hidden layer to one and compared the neural networks with different numbers of hidden nodes in the hidden layer. Figure 5 shows the EERs of the utterance type recognizer with one hidden layer and various numbers of hidden nodes when the order of the polynomial expansion is fixed to be three. The recognizer with a linear perceptron, which is equivalent to a multilayer perceptron with one hidden node, achieved the lowest EER. In this open test, the larger the number of hidden nodes, the larger the EER of the recognizer. The larger number of hidden nodes used in the neural networks means that the neural networks will be more specific in learning the training set. Since the training set does not cover all variation of the pitch contour of the final word of the speech data in the testing set, the more specific the neural networks learn the training set, the larger the error of the neural networks in the recognition of the utterance type of the testing set.

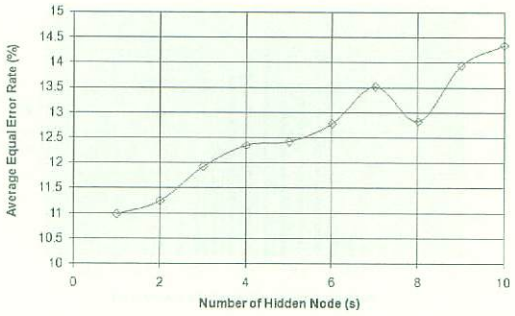

Fig. 5 Equal Error Rate of the open test of the utterance type recognizer using the third order polynomial expansion, one hidden layer and various numbers of hidden nodes.

Table 6 The EER of the open test of the utterance type recognizer using various orders of the polynomial expansion and a linear perceptron.

Order of the polynomial expansion	EER (%)
1	17.2
2	15.8
3	10.9
4	11.1
5	38.3

Next, we evaluated the utterance type recognizers using different orders of the polynomial expansion, where the linear perceptron was used as the classifier. Table 6 shows the EERs of the utterance type recognizer when the order of polynomial expansion is varied. The lowest EER was achieved when the third order polynomial expansion was used. Further increase of the order of the polynomial expansion increased the EER.

Finally, we compared the EERs of the proposed automatic utterance type recognizer with the EERs when the segmentation of the final words was carried out manually. The order of polynomial expansion is fixed to three. The recognizer with the manual segmentation achieved 88.1% accuracy, which is 1.0 point worse than the automatic recognizer did. The average errors in the estimation of the beginning time and the duration of the final word are 43.5 ms and 75.4 ms, respectively. This small difference made the recognition rates of both the utterance type recognizers comparable.

The average error rate of the proposed utterance type recognizer was higher than that of the utterance type recognizer based on Fujisaki model [13]. This is because the recognizer in [13] is speaker-dependent and evaluated

using the testing set that included in the training set. On the other hand, the proposed recognizer is speaker-independent and evaluated using the testing set that differs from the training set in both the sentence and the speaker. It is expected to be more robust in real application. The equal error rate obtained by our method is 10.9% which is about five times larger than the error rate 2.2% obtained in the perceptual test. It was reported that automatic continuous speech recognition was more than ten times erroneous than human performance [33]. While the difficulties of the tasks are largely different, we believe that our method can be applicable to real use.

6. CONCLUSIONS

This paper reports our study of an automatic utterance type recognizer to identify DQs and statements in Indonesian speech. The findings of the study confirmed that the use of the *final word* of the utterance and the pitch contour information was effective in classifying Indonesian DQs and statements. The highest recognition rate 89.1% was achieved using the third order polynomial expansion as the feature extractor and a perceptron as the classifier. When using *final words* that contain only two syllables (*final-two-syllables*), the recognition rate is 0.1 point higher than using all *final words*. The proposed automatic recognizer can be combined with another automatic system such as a speech recognizer to build a larger automatic spoken system such as a spoken dialogue system or a spoken understanding system.

In the future, we plan to develop a new larger speech database, especially from female speakers in order to cover larger variation of the pitch contour of the utterances and to design an automatic utterance type recognizer that covers all types of utterances. The automatic segmentation module of the recognizer will be further developed following the development of the speech database. Since our method is based on statistical approach, it can be easily applicable for other languages once the utterance part to be used for classification is specified. We plan to apply our method to utterance type recognition of other similar languages such as Malay spoken in southern Thailand, Malaysia, and Brunei and Tagalog spoken in the Philippines.

ACKNOWLEDGMENTS

The authors would like to thank to the AUN/SEED Net for the scholarship of doctoral degree program to the first author.

REFERENCES

[1] M. Fishel, "Dialogue act recognition techniques," in GSLT/NGSLT course on dialogue systems: Linkoping University, Sweden (2006).

[2] B. Carpenter and J. Chu-Carrol, "Spoken dialogue systems," Lucent Technologies, Bell Labs Innovations (1999).

[3] H. Wright, "Automatic utterance type detection using suprasegmental features," *5th Int. Conf. Spoken Language Processing (ICSLP 98)*, Sydney, Australia (1998).

[4] H. Wright, M. Poesio and S. Isard, "Using high level dialogue information for dialogue act recognition using prosodic features," *ESCA Tutorial and Research Workshop on Dialogue and Prosody*, Eindhoven, The Netherland (1999).

[5] S. Grau, E. Sanchis, M. J. Castro and D. Vilar, "Dialogue act classification using a Bayesian approach," *9th Int. Conf. Speech and Computer (SPECOM 2004)* (2004).

[6] A. G. Adam, *et al.*, "Modeling prosodic dynamics for speaker recognition," *ICASSP 2003*, Hongkong (2003).

[7] H. Alshawi, "Effective utterance classification with unsupervised phonotactic models," *2003 Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, Edmonton, Canada (2003).

[8] W. Wahlster, T. Bub and A. Waibel, "Verbmobil: The combination of deep and shallow processing for spontaneous speech translation," *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Germany (1997).

[9] J. Lee, G. C. Kim and J. Seo, "A dialogue analysis model with statistical speech act processing for dialogue machine translation," *Spoken Language Translations EACL '97 Workshop*, Budapest, Hungary (1997).

[10] N. Reithinger, R. Engel, M. Kipp and M. Klesen, "Predicting dialogue acts for a speech-to-speech translation system," *ICSLP 96* (1996).

[11] V. J. van Heuven and E. van Zenten, "Speech rate as a secondary prosodic characteristic of polarity questions in three languages," *Speech Commun.*, **47**, 87–99 (2005).

[12] A. Halim, *Intonation in Relation to Syntax in Indonesian* (Australian National University, Canberra, 1981).

[13] N. Effendy, E. Maneenoi, P. Charnvivit and S. Jitapunkul, "Intonation recognition for indonesian speech based on Fujisaki model," *Interspeech 2004*, Korea (2004).

[14] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, 2nd ed., rev. and expanded ed. (Mercel Dekker, Inc., New York, 2001).

[15] M. Akagi and T. Ienaga, "Speaker individualities in fundamental frequency contours and its control," *EuroSpeech '95*, Madrid, Spain (1995).

[16] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Commun.*, **16**, 165–173 (1995).

[17] C. Odé, "On the perception of prominence in Indonesian," in *Semaian 9: Experimental Studies of Indonesian Prosody*, C. Odé and V. J. van Heuven, Eds. (Department of Languages and Cultures of South-East Asia and Oceania, Leiden University, Leiden, 1994), pp. 27–107.

[18] E. van Zanten and V. J. van Heuven, "Word stress in Indonesian: Its communicative relevance," *Bijdragen tot de Taal-, Land- en Volkenkunde*, **154**, 129–149 (1998).

[19] Samsuri, *Analisa Bahasa: Memahami Bahasa Secara Ilmiah* (Erlangga, Ltd., Jakarta, 1978).

[20] N. Reithinger and E. Maier, "Utilizing statistical dialogue act processing in verbmobil," *33rd Annu. Meet. Association for Computational Linguistics* (1995).

[21] K. Ries, "HMM and neural network based speech act detection," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, Arizona (1999).

[22] S. Keizer, R. op den Akker and A. Nijholt, "Dialogue act recognition with Bayesian networks for Dutch dialogues," *3rd SIGdial Workshop on Discourse and Dialogue*, Philadelphia, PA, USA (2002).

[23] K. Samuel, S. Carberry and K. Vijay-Shanker, "Automatically selecting useful phrases for dialogue act tagging," *4th Conf. Pacific Association for Computational Linguistics (PACLING '99)*, Waterloo, Ontario, Canada (1999).
[24] C. T. Ishi, "Perceptually-related F0 parameters for automatic classification of phrase final tones," *IEICE Trans. Inf. Syst.*, **E88-D**, 481–488 (2005).
[25] L. Levin, C. Langley, A. Lavie, D. Gates, D. Wallace and K. Peterson, "Domain specific speech acts for spoken language translation," *4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan (2003).
[26] Q. Yan, S. Vaseghi, D. Rentzos, C.-H. Ho and E. Turajlic, "Analysis of acoustic correlates of British, Australian and American accents," *ASRU 2003* (2003).
[27] D. Bolinger, *Intonation and Its Parts: Melody in Spoken English* (Stanford University Press, Palo Alto, CA, 1986).
[28] P. J. Watson and D. Hughes, "The relationship of vocal loudness manipulation to prosodic F0 and durational variables in healthy adults," *J. Speech Lang. Hear. Res.*, **49**, 636–644 (2006).
[29] H. Levitt and L. R. Rabiner, "Analysis of fundamental frequency contour in speech," *J. Acoust. Soc. Am.*, **49**, 569–582 (1971).
[30] S. H. Hwang and S. H. Chen, "Neural-network-based F0 text-to-speech synthesizer for Mandarin," in *IEE Proc. Vision Image Signal Process.*, **141**, 384–390 (1994).
[31] W. C. Yip and D. L. Barron, "Speech recognition using polynomial expansion and hidden Markov models," in http://www.patentstorm.us/patents/6928409.html. US: Freescale Semiconductor, Inc. (2005).
[32] J. Yuan, C. Shih and G. P. Kochanski, "Comparison of declarative and interrogative intonation in Chinese," *Speech Prosody 2002*, Aix-en-Provence, France (2002).
[33] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The Hidden Markov Model Toolkit [HTK] version 3.2.1 for Speech Processing* (Cambridge University Engineering Department, 2002).
[34] Entropic, *ESPS Quick Reference* (Entropic Research Laboratory, Inc., Washington, D.C., 1998).
[35] Entropic, *ESPS Version 5.0 Programs Manual* (Entropic Research Laboratory, Washington, D.C., 1993).
[36] S. Katagiri, *Handbook of Neural Networks for Speech Processing* (Artech House, Boston, 2000).

Nazrul Effendy received the B.Eng. degree in instrumentation technology of nuclear engineering and M.Eng. degree in Electrical Engineering from Gadjah Mada University, Indonesia in 1998 and 2001, respectively. He has been a faculty staff at the Department of Engineering Physics, Gadjah Mada University since 1998. Since November 2003, He has been doing a Ph.D. sandwich program between Chulalongkorn University, Thailand and Tokyo Institute of Technology, Japan. His research interests include speech signal processing, prosody, and pattern recognition. He was a member of ISCA during 2004–2006.

Koichi Shinoda received the B.S. degree in 1987, M.S. degree in 1989, both in physics from the University of Tokyo, and Dr. Eng. degree in computer science from Tokyo Institute of Technology in 2001. He is currently an Associate Professor with Tokyo Institute of Technology. His research interests include speech recognition, statistical pattern recognition, and human interface. He received the Awaya Prize from the Acoustic Society of Japan in 1997 and the Excellent Paper Award from the Institute of Electronics, Information, and Communication Engineers in 1998. He is a member of IEEE, ASJ and IEICE.

Sadaoki Furui received B.S., M.S., and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University, Tokyo, Japan in 1968, 1970, and 1978, respectively. He has authored or coauthored over 700 published articles. He is currently a Professor and Head of the Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, an Instructor at the University of Tokyo, a visiting researcher at NHK (Japan Broadcasting Corporation) and a visiting researcher at National Language Research Institute, Japan. He is a Fellow of IEEE, ASJ, and IEICE. He served as President of PC-ICSLP (2000–2004), ISCA (2001–2005), and ASJ (2001–2003). He served as an Editor-in-Chief of the Journal of Speech Communication (1997–2001), Chief Editor of the Journal of the ASJ (1997–1999), and Chief Editor of the English Journal of IEICE (2001–2003). He is now serving as an Editorial Board member of the Journal of Computer Speech and Language and the Journal of Speech Communication. He is also serving as a Board member of the IEICE.

Somchai Jitapunkul received the B.Eng. and M.Eng. degrees in Electrical Engineering in 1972 and 1974, respectively from Chulalongkorn University, Thailand. In 1976 and 1978, he received the D.E.A. and Dr. Ing. degrees, respectively, in "Signaux et Systems Spatio–Temporels" from Aix–Marseille University, France. He was appointed as a lecturer in the department of Electrical Engineering at Chulalongkorn University in 1972, Assistant Professor in 1980, and Associate Professor in 1983. In 1993, he was the founder of Digital Signal Processing Research Laboratory where he became the head of this laboratory from 1993 to 1997. From 1997 to 1999 and 1999 to 2003, he was appointed as the head of Communication Division and the head of the Department, respectively. He also held the position of Associate Dean for Information Technology, Faculty of Engineering from 1993 to 1995. His current research interests are in image and video processing, speech and character recognition, signal compression, DSP in telecommunication, software defined radio, smart antenna, and medical signal processing. He is a member of IEICE.