IMPROVE SEGMENTAL EVALUATION USING SYLLABIC FILLERS IN INDONESIAN ISOLATED WORD RECOGNITION

Linda Indrayanti[1], Yoshifumi Chisaki[1], and Tsuyoshi Usagawa[1]

[1]Graduate School of Science and Technology (GSST), Kumamoto Univ., Kumamoto 860-8555

{linda@hicc., chisaki@, and tuie@}cs.kumamoto-u.ac.jp

ABSTRACT

In the application of utterance training system, an intensive evaluation from segmental aspect is essential. Feedback provided is needed for further remedial process in learning so that it is less fair to refer to recognition result only in measuring non-native speech. Utilizing keyword spotting applied for isolated word recognition, performance of sub-word uttered by non-native is measured from its correct detection and false alarm rate. A critical issue in utterance training is non-keywords yielded from dis-fluency of non-native speech, which are discussed in this study. The keywords are represented by context-dependent HMMs trained using pooled data from native and non-native, and the non-keywords are represented by acoustic model with filler trained using syllabic other than the keywords. Using small data the achieved improvement is 63% compared to 43% when the filler is not used.

1. INTRODUCTION

In some applications, the performance can be improved by implementing a keyword spotting technique, compared with the case when the syntax contains only keywords without non-keywords model. The task of keyword spotting in this study is to detect a set of sub-word in the input isolated word of utterance training system. In such application, the input speech includes out-of-task sub-words caused by dis-fluency of non-native speech. By introducing a keyword spotting technique, these phenomena can be covered, and therefore the system gives non-native flexibility to speak naturally and less discouraging. The performance of an HMM-based keyword spotter depends on the ability of the filler models to represent non-keyword without rejecting the correct keywords (false alarms). Therefore, the choice of an appropriate filler model set is a critical issue. This study proposes a method for modelling the non-keyword based on the use of sharing parameter in context-dependent HMMs based. Using extraneous information occurred at the front and the end of non-native speech other than the keyword one to train filler models, the objective is to develop evaluation procedure for non-native speech, and to overcome the problem of the high rate of false alarms.

2. DATA COLLECTION

The speech material, an available collected corpus is a set of Indonesian basic words. The data were collected for simple dictation system taken from frequently used words in daily life. Non-native database consists of four set data read normally by eight male non-native speakers from eight different nationalities (see Table 1), recorded over DAT and down-sampling with frequency 16 kHz. They are in the same level of proficiency (beginner level) in Indonesian language. They never had any experiences in Indonesian language before this

Table 1. Non-native speakers.

ID	Gender	L1s
AR	M	Arabic
CH	M	Chinese
EN	M	English
FR	M	France
GR	M	German
JP	M	Japanese
SP	M	Spanish
TR	M	Turkish

インドネシア語孤立単語認識における音節フィーラを用いたセグメンタル評価の改良

リンダ　インドラヤンティ, 苣木 禎史, 宇佐川 毅 (熊本大学)

experiment. Practice under native guidance was given briefly just before recording task that was set up under identical condition in anechoic room. As error occurred during the process, they were required to retake for the mistake only.

3. BASELINE SYSTEM

Viterbi algorithm recognized test set against acoustic model and outputted phone transcription. Then it would be compared with the correct transcription from forced alignment process. Keyword spotting is carried out to compute figure of merit (FOM) [1]. The keywords include 35 phonemes (vowels inclusion diphthongs and consonants. The FOM is calculated by analyzing list of detected keywords in order. From top order of the list, detection is counted for each *false alarm, FA*; the number of correct phones accepted by the sys tem as a percentage of the total number of phones.

3.1. Native Gender Independent (NGI) System

Native Gender Independent system (NGI) system consisted with one-hundred words uttered by 21 females and 21 males, in total 4200 native utterances used for training purpose. Acoustic models were context-dependent linear no-skip 3-state triphone HMMs clustered via tree-based clustering. Each state used single Gaussian mixture components. The speech features used comprise 12 MFCC coefficients and corresponding Δs and energy coefficients. The acoustic models, set-up on isolated word with IPA transcription, were trained by means of forward-backward estimation on isolated word training set. Word error rate (WER) on native is 15% and on non-native is 81%.

3.2. Non-Native Speaker Dependent (NNSD) System

Non-Native Speaker Dependent system (NNSD) acoustic models were trained on non-native training set described in data collection section. Measurement, labelling and training procedure were kept the same as those used to train the NGI system. As expected, the NNSD system performed better on non-native than on the NGI system. This result provided some assurance that the NNSD acoustic models were trained adequately. WER on non-native is 11% and on native is 51.3%.

3.3. Pooled Native Gender Independent - Non Native Speaker Dependent (NGI-NNSD) System

In this experiment, native and non-native training data were pooled together to build acoustic models. From speaker population, native data were much more in variety than non-native data. In total, there were 4200 natives and 3200 non-natives utterances for training purpose. When testing on non-native, the NGI-NNSD system gets WER of 18.8%, and it gives 13.8% of WER tested on native. Male dominance in non-native speech might give contribution to this reason. It is found that the NGI-NNSD system is the most flexible system compare to the two others. The NGI system was too strict by means of too many rejections made that could be discouraging for non-natives. The NNSD system was tailored to fit only for eight non-natives that would give so much adaptation to some specific mother tongues. In between, the NGI+NNSD system is able to define relative confidence between native and non-native speech as proved on related WER.

3.4. Experiments and Results

Inclusion of non-native in the NGI+NNSD system does improve the performance. Tying states mechanism within context-dependent phone sets for clustering similar acoustic parameter to make robust estimation may not work effectively. A possible reason for this is insufficiency of non-native data thus less reliable in estimating statistics of non-native speech. Moreover, tying mechanism using context decision tree was built from native speech only to model the context of non-native speech. In result, the decision tree does not represent the context of the non-native speech very accurately. Figure 1 shows the performance of context-dependent phone across L1s.

The NGI-NNSD system leads to improve performance even not exceeding the NNSD system. This is happened by some reasons: high proportion of dis-fluency and characterristics of non-native speech, only one speaker in one group of L1 and reference labels were obtained by forced-alignment based on phone-level transcript rather than by careful hand-la-

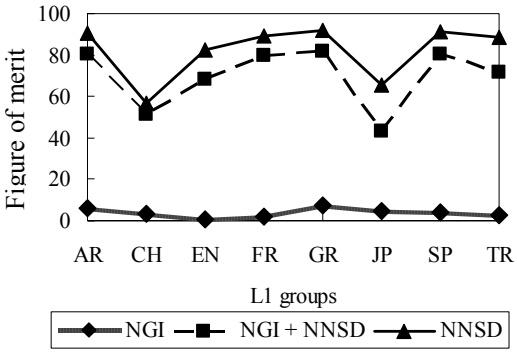

Fig. 1. FOM across three-acoustic model settings.

Table 2. Phone error rate of three-acoustic model settings.

	Acoustic Model (% Accuracy)		
	NGI	NNSD	NGI -NNSD
Vowels	60.3	12.2	30.2
Consonants	72.2	19.9	39.0
Mean	**66.3**	**16.0**	**34.6**

belling. Table 2 shows that consonants are mispronounced more often than vowels and diphthongs.

3.5. ASR versus Human Evaluation

The standard way of evaluating the performance of automatic alignment is to compare the results with manually evaluated one. The recordings were presented via headphones to two evaluators who were asked to assess the performance of the speakers by overall speaker efficiency, accepted or not accepted. Faulty or unusual pronunciations occurred in the utterances were carefully scrutinised to determine which phone had caused an error. The findings were then, examined and the error noted. Approximately prediction by human judgment, there are 9 *FAs* found in comparing evaluation result between human and ASR system.

Conversely, there is a cognitive cost for undetected miscue, i.e., word misread or omitted by non-native but accepted by ASR. Some words may exactly have similar composition in phoneme and merely be differed with one phone at the beginning/at the end as in Table 3.

Table 3. Test data set with the word spoken and confusable set that may caused improper pronunciation.

Words spoken	Confusable words
Enak	Anak
Hutang	Hutan
Jamur	Campur
Kamu	Kamus
Kacang	Datang

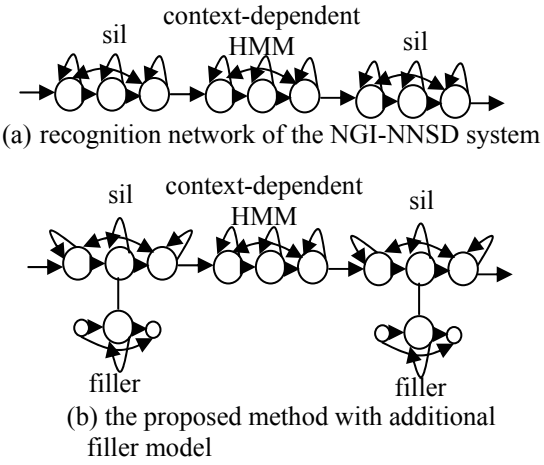

Fig. 2. Keyword spotting model: (a) Recognition network without filler (b) Keyword spotting network with filler model.

This condition can lead to problem when non-native is not able to make distinction for these words.

4. CONTEXT-DEPENDENT PHONE WITH FILLER MODEL

As listed in Table 3, mispronunciation may caused by pronunciation similarity of words. To handle this, simple *filler* model was utilized to take care of improper pronunciation appeared at the beginning and the end of word. Out-of-vocabulary phones as result from mispronunciation are treated as extra information and are categorized as part of the train ing data. The proposed method was composed of the following components:

- Representation of keyword uses the NGI-NNSD acoustic model.
- Filler refers to additional word in the vocabulary to dis-fluency of the front and the end of isolated words.
- As Fig. 2a and 2b adopted from HTK[1], modification is done with the *sp* (short-pause) model that originally represented *pause* between words in continuous speech application. *Filler* model is built as the same mechanism as *sp* model but to be concatenated to context-dependent HMMs at the beginning and the end of each word. Parallel line from *sil* to *filler* means sharing parameter between two models.

4.1 Experiments and Results

The same number of non-native as listed in Table 1 gets involved in this experiment in order to evaluate the proposed method. They

as listed in Table 3 are required to utter five words twice, one set (40 utterances) for training and the other one is for testing. Two human evaluators measure whole performance of each test dataset and give positive mark for the accepted one. Two ASR systems (ASR without and with filler model) work in recognition rate accuracy referred to the template transcription. Figure 3 shows the recognition performance of the test dataset where only 70% of test words are correctly pronounced as shown by score of human evaluator. In comparing two systems, the ASR with *filler* model has better accuracy compare to ASR without *filler* model. It observes how effective the filler model worked. In other word, the proposed method works as a threshold between human evaluators and the ASR systems. Figure 4 shows *Receiver Operating Curves* (ROC) to quantify the accuracy of confidence score from the ASR system. The ROC depicts the *hit* rate (the number of semantically correct phones accepted by the system as a percentage of the total number of phones), as a function of the *FA* rate. A perfect system would have its *hit*/*FA* curve follow the left vertical axis (0% FA) and the top horizontal axis (100% hit). In average, the ASR system with *filler* model leads to 5.3% improvement in the *hit* rate or conversely a 6.2% reduction in the *FA* rate compare to the ASR without *filler* model.

CONCLUSIONS

The results have identified that sharing data between native and non-native has potential for significantly providing confidence assessment both for native and non-native speech. They also demonstrate that the proposed method outperforms the baseline system in out-of-vocabulary phone as a result from pronunciation similarity. With the small set of test data, it outperforms with a 23 points gain of recognition rates over the baseline system. In further study, development of background acoustic model will be investigated.

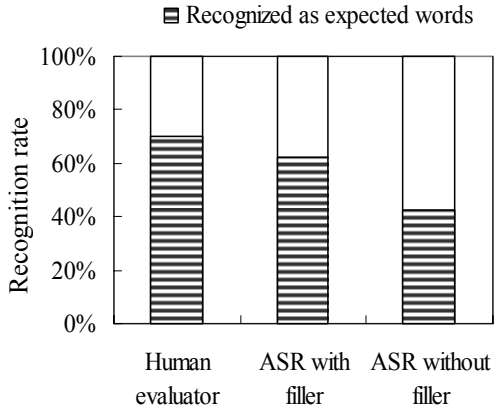

Fig. 3. Evaluation by human evaluators, ASR with filler model, and ASR without filler systems.

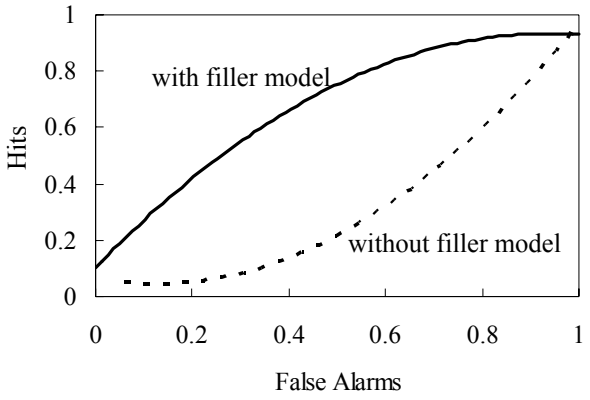

Fig. 4. The ROC curve comparing performance of the baseline system without filler model and the keyword spotting with filler model.

REFERENCES

[1] HTK software package Ver.3.3.

[2] L. Indrayanti, T. Usagawa, Y. Chisaki, T. Dutono, "On Segmental Errors in Indonesian Language as A Second Language," in Proc. of the 8th ITHET, Kumamoto, July 07.

[3] L. Indrayanti, T. Usagawa, Y. Chisaki, "Indonesian Isolated Word Keyword Spotting Based on Syllabic Fillers", in *Proc. of ASJ Autumn, Yamanashi, Sept07.*

[4] P. Heracleous, and T. Shimizu, "A Novel Approach for Modelling Non-Keyword Intervals in a Keyword Spotter Exploitting Acoustic Similarities of Languages," Speech Comm. 45 (2005) 373-386.

[5] A. Neri, C. Cucchiarini, H. Stirk, "Segmental Errors in Dutch as a second language: how to establish priorities for CAPT", *in Proc. of InSTIL/ICALL,Venice, 2004.*

[6] A. Gunawardana, H. Hon, and L.Jiang, "Word-Based Acoustic Confidence Measures for Large-Vocabulary Speech Recognition," Microsoft Research, Redmond USA.

[7] L. Hamel, "Model Assessment with ROC Curves," Dept. of Comp. Scie. & Stat. USA.