Journal of Computer Science 6 (11): 1334-1340, 2010
ISSN 1549-3636
© 2010 Science Publications

Acoustic Model Adaptation for Indonesian Language Utterance Training System

Linda Indrayanti, Yoshifumi Chisaki and Tsuyoshi Usagawa
Department of Computer Science and Electrical Engineering,
Graduate School of Science and Technology,
Kumamoto University,
2-39-1 Kurokami Kumamoto, 860-8555, Japan

Abstract: Problem statement: In order to build an utterance training system for Indonesian language, a speech recognition system designed for Indonesian is necessary. However, the system hardly works well due to the pronunciation variants of non-native utterances may lead to substitution/deletion error. This research investigated the pronunciation variant and proposes acoustic model adaptation to improve performance of the system. **Approach:** The proposed acoustic model adaptation worked in three steps: to analyze pronunciation variant with knowledge-based and data-derived methods; to align knowledge-based and data-derived results in order to list frequently mispronounced phones with their variants; to perform a state-clustering procedure with the list obtained from the second step. Further, three Speaker Adaptation (SA) techniques were used in combination with the acoustic model adaptation and they are compared each other. In order to evaluate and tune the adaptation techniques, perceptual-based evaluation by three human raters is performed to obtain the "true" recognition results. **Results:** The proposed method achieved an average gain in Hit + Rejection (the percentage of correctly accepted and correctly rejected utterances by the system as the human raters do) of 2.9 points and 2 points for native and non-native subjects, respectively, when compared with the system without adaptation. Average gains of 12.7 and 6.2 points for native and non-native students in Hit + Rejection were obtained by combining SA to the acoustic model adaptation. **Conclusion/Recommendations:** Performance evaluation of the adapted system demonstrated that the proposed acoustic model adaptation can improve Hit even though there is a slight increase of False Alarm (FA, the percentage of incorrectly accepted utterances by the system of which the human raters reject). The performance of the proposed acoustic model adaptation depends strongly on the effectiveness of state-clustering procedure to recover only in-vocabulary words. For future research, a confidence measure to discriminate between in-vocabulary and out-vocabulary words will be investigated.

Key words: Utterance training system, Indonesian language, acoustic model adaptation, perceptual-based evaluation

INTRODUCTION

In recent years, there is an increase interest of foreign students to study in Indonesia especially on Indonesian language and local culture. Due to limited time of their study (range: from 1 month to 1 year), it would be very beneficial for them to study Indonesian language preliminary so that their study time becomes more effective. Started from this condition, an initial idea to develop an Utterance Training System (UTS) for Indonesian language came up. In addition, speaking practice is necessary skill to complement reading and listening lessons, which are available from various books and educational software. The current isolated word recognizer for Indonesian language, called as the baseline system here after which was trained by native utterances data will encounter drastic degradation on non-native utterances. The reason is non-native subjects often make substitution or deletion error due to pronunciation variant contained in their utterances. Therefore, adaptation to compensate non-native pronunciation variants is required in order to improve recognition accuracy of the baseline system. This study studies an acoustic model adaptation in the Indonesia

Corresponding Author: Linda Indrayanti, Department of Computer Science and Electrical Engineering, Graduate School of Science and Technology, Kumamoto University, 2-39-1 Kurokami Kumamoto, 860-8555, Japan

language UTS based on non-native utterances. Various approaches had been proposed regarding with an acoustic model adaptation on non-native utterances, for example: to use characteristics of the mother tongue (source language) of non-native subject in the evaluation of his/her pronunciation (Moustroufas and Digalakis, 2006). A dictionary modification, an acoustic model adaptation and manipulation were typical techniques, which could improve non-native utterances as shown in (Oh *et al*., 2007; Alia and Al Mograbi, 2007). In line with the main idea of some published works, the proposed acoustic model adaptation works as follows: frequently mispronounced phones with their pronunciation variants of non-native subjects are analyzed by performing alignment analysis between knowledge-based and data-derived results. Knowledge-based method utilizes human raters to carry out phonetic analysis between Indonesian language and non-native language. On the other hand, data-derived method utilizes the system to align automatically non-native utterances with reference transcription of correct utterance and creates monophone-based confusion matrix. Result from the alignment analysis is a list of mispronounced phones with their variants, which is used to perform an acoustic model adaptation on a state-clustering procedure. Presence of human raters in the proposed acoustic model adaptation is necessary in order to provide a standard evaluation against recognition results of the system, as mentioned in (Neumeyer *et al*., 1996; Franco *et al*., 1997). Perceptual-based evaluation of human raters is not only to simply value non-native utterances as accepted/rejected but also to analyze and locate specific errors on segmental aspects. Further, the acoustic model adaptation is combined with three speaker adaptation techniques Maximum Likelihood Linear Regression (MLLR) as proposed in (Goronzy *et al*., 2004; Giuliani *et al*., 2006; Haraty and El Ariss, 2007), Constrained MLLR (CMLLR) and Vocal Track Length ormalization (VTLN) as proposed in (Hariharan *et al*., 2002; Sundermann *et al*., 2003; Legetter and Woodland, 1995; Shen and Reynolds, 2008; Al-Haddad *et al*., 2009; Gales and Young, 2008) in order to eliminate inter-speaker variability. Performance of the proposed acoustic model adaptation is evaluated in five measures of alignment analysis between recognition results and perceptual based evaluation: Hit, False Alarm (FA), Miss, Rejection and Hit + Rejection.

MATERIALS AND METHODS

Speech database: Speech databases are constructed to be used for training and testing purposes. The material is composed of 100 isolated words, which are used to develop a native and a non-native speech database. The data (frequently used every day words) are collected for simple isolated word recognition.

The native speech corpus consists of utterances from 42 native speakers, most of which of Javanese mother tongues (Tan and Hussain, 2009). Each word normally uttered twice. From 8400 native utterances, 4200 (50%) of them were used for training. The other native speech database consists with 10 native speakers (1000 utterances) is developed to be used in evaluating the performance of the recognizer.

The non-native speech database consists of utterances from 8 males and 1 female student. Those non-native students have no experience in learning Indonesian language before this experiment (in other words, they are all at the same beginner level). A brief explanation and pronunciation practice under native guidance is given just before recordings take place. Non-native students utter each word normally four times. Once a mispronunciation occurs during the process, they are required to redo the task to correct the mistake only. From the 3600 non-native utterances, 1800 (50%) are used for training purpose. Another non-native speech database contained with 4 males-1 female students (500 utterances) is developed for testing.

Acoustic model adaptation: An Acoustic Model Adaptation (AMA) method is proposed in order to improve recognition performance of the baseline system evaluated on non-native utterances. The proposed adaptation method consists of three steps:

1. To observe pronunciation variant made by non-native students in Indonesian language with two different ways: knowledge-based and data-derived methods (Wester, 2003)

Knowledge-based method uses general knowledge about Indonesian language and non-native languages and the procedure is as follows:

- Three human raters (Indonesian graduate students whose major are engineering) are equipped with headphones, recorded speech from 5 non-native students, the list of 100 words with transcriptions and 5 lists of foreign phonology classification. Brief explanation on how to perform the evaluation is provided beforehand. Each rater is accompanied by one of authors during evaluation task to keep a steady performance measure
- In response, human raters evaluate each utterance based on segmental quality. Any unusual pronunciation is noted and carefully scrutinized to find its error

- To test reliability of each human rater, evaluations on the same set utterances are carried out twice for each human rater. And it is found that intra-rater reliability is 0.89. The degree of agreement among human raters (inter-rater reliability) is also high about 0.93
- Output from this process: one human rater has 5 evaluation results of 5 non-native students. The judgment of each unusual pronunciation by the three human raters is evaluated by majority rule to lead a decision. When two human raters agree to accept a certain pronunciation while one human rater rejects it, voting is carried out to determine the result. As a result, one list of mispronounced phones with their pronunciation variants summarized of 5 non-native students is obtained

Data-derived method uses the baseline system which is trained by pooled data between native and non-native utterances, to perform automatic alignment of non-native utterances with reference transcription of correct utterance and to output monophone-based confusion matrix. Confusion matrix consists with a number of phones, which are correctly classified as the same phone or incorrectly classified as another phones.

2. To carry out alignment analysis of knowledge-based and data-derived results. Three human raters work collaboratively to align the list of mispronounced phones obtained by knowledge-based with the frequently mispronounced phones obtained by confusion matrix. As a result, list of frequently mispronounced phones with their corresponding pronunciation variants are obtained as shown in Table 1
3. To perform a state-clustering procedure based on the results shown in Table 1. The state-clustering of the proposed acoustic model works as follows:

- An initial set of a 3 state left-right monophone model is created and trained with native and non-native utterances
- A set of context-dependent triphone models is made by cloning monophone models
- In a conventional state-clustering, for each set of context-dependent triphone derived from the same monophone, corresponding states were clustered. For example (triphone l-c+r), clustering was performed for each center monophone /c/ in the triphone-based acoustic models and all corresponding left-right phones were tied to /c/. However, in the state-clustering of the proposed acoustic model adaptation method, clustering is performed in two conditional ways: for the center monophones with pronunciation variants and the other is for those without. If the center monophones of /c/ has a pronunciation variant /c'/, as a result from the alignment analysis of perceptual-based and system evaluation on non-native utterances, the center monophone /c/ or /c'/ in the triphone-based acoustic models are pooled together and the corresponding left-right monophones are clustered. Otherwise, the conventional state-clustering is performed
- The number of mixture components in each state is incremented and the models re-estimated until the best performance reached

AMA in combination with speaker adaptation: A UTS should be speaker independent i.e., inter-speaker variability should be eliminated. Various adaptation methods have been used to deal with inter-speaker variability. An approach used to solve this problem is to use a speaker adaptive training to deal with inter-speaker variability. The main idea is to normalize the speech signal of a new utterance such that it is similar to the average utterances. Another way is a parameter adaptation. A transformation is used to minimize the mismatch between new utterances and average utterances.

This study shows simple and commonly used speaker adaptation techniques (MLLR (Goronzy *et al*., 2004; Giuliani *et al*., 2006), CMLLR (Hariharan *et al*., 2002; Sundermann *et al*., 2003; Legetter and Woodland, 1995; Shen and Reynolds, 2008) and VTLN) to compensate for speaker-specific differences caused by non-native language influence for isolated words. Table 2 shows the results of the baseline system adapted with speaker adaptation techniques (MLLR, CMLLR and VTLN). And Table 3 shows the results of the baseline system adapted with the combination of AMA and speaker adaptation techniques (MLLR, CMLLR and VTLN).

Assessment and evaluation:

Automatic analysis: The HTK Tools package (Woodland *et al*., 1994) is used for speech analysis, acoustic model training and speech recognition purposes.

Table 1: A list of frequently mispronounced phones with their corresponding pronunciation variants as a result of alignment analysis between data-derived and knowledge-based methods

Target phones	Pronunciation variants
Vowel	
/ê/	/e/
Consonants	
/c/	/k/
/l/	/r/
/v/	/b/

Table 2: Results of alignment analysis between recognition results and perceptual-based evaluation for native and non-native utterances evaluated on the baseline system adapted with three Speaker Adaptation (SA) techniques (MLLR, CMLLR and VTLN)

		Alignment analysis (%)				
Systems	Subjects	Hit	FA	Miss	Rejection	Hit + Rejection
The baseline system	Non-native	60.4	15.4	16.6	7.6	68.0
	Native	86.2	0.0	13.8	0.0	86.2
The baseline system + MLLR	Non-native	67.1	18.0	9.9	5.0	72.1
	Native	97.3	0.0	2.7	0.0	97.3
The baseline system + CMLLR	Non-native	68.3	17.6	8.7	5.4	73.7
	Native	96.1	0.0	3.9	0.0	96.1
The baseline system + VTLN	Non-native	65.3	17.6	11.7	5.4	70.7
	Native	97.4	0.0	2.6	0.0	97.4

Table 3: Results of alignment analysis between recognition results and perceptual-based evaluation for native and non-native utterances evaluated on the baseline system with Acoustic Model Adaptation (AMA) in combination with three Speaker Adaptation (SA) techniques (MLLR, CMLLR and VTLN).

		Alignment analysis (%)				
Systems	Subjects	Hit	FA	Miss	Rejection	Hit + Rejection
The baseline system + AMA	Non-native	64.0	17	13.0	6	70.0
	Native	89.1	0	10.9	0	89.1
The baseline system + AMA + MLLR	Non-native	70.9	19	6.1	4	74.9
	Native	99.0	0	1.0	0	99.0
The baseline system + AMA + CMLLR	Non-native	70.8	19	6.2	4	74.8
	Native	98.7	0	1.3	0	98.7
The baseline system + AMA + VTLN	Non-native	68.9	19	8.1	4	72.9
	Native	98.9	0	1.1	0	98.9

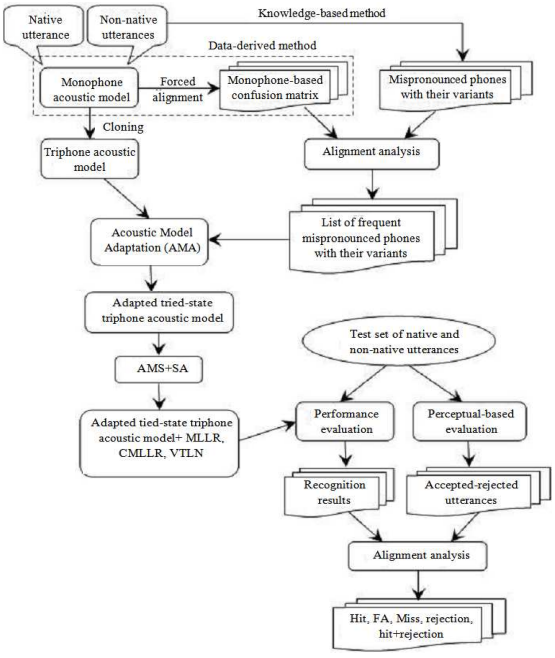

Fig. 1: A block diagram of the proposed acoustic model adaptation based on pronunciation variant of non-native utterances obtained from alignment analysis between knowledge-based and data-derived methods

There are independent programs for each step of training and recognition processes. A set of phoneme level HMMs is trained on the utterances (and the labels) in the training set. During the training process, each utterance is encoded and the relevant features are extracted based on the choice of features, window size and frame period. Each HMM state is modeled initially by a mixture-of-Gaussians of size 1 and trained using four-cycles of the Baum-Welch re-estimation. This is repeated until the best performance was reached. After obtaining the phoneme level HMMs, testing process is conducted by applying these HMMs to the test set using forced alignment and the Viterbi algorithm. The testing process generates a set of auto-labeled phones (phone name, start and end time) for each utterance. The recognition performance of the system is calculated by counting the correctly recognized words. The overall process including adaptation procedure is drawn in Fig. 1.

Alignment analysis between recognition result and perceptual-based evaluation: Human raters take part as a standard evaluation in evaluating non-native utterances against recognition results of the system. Perceptual-based evaluation obtained by human raters should be target of the system in measuring its performance reliability. Human raters evaluated the

quality of each non-native utterance for its entire content (overall pronunciation) as follows:

- Three human raters as previously mentioned are used. They work voluntary for this task that takes about 3 h. A brief explanation on how to perform the evaluation on non-native utterance is performed before the task. Each rater is provided with the list of 100 words with transcriptions. In total there are 5 lists of non-natives students to be evaluated by each rater
- Non-native utterances are presented via headphones to human raters who are asked to assess the performance of each non-native utterance. All raters listen to the speech material and perform their own evaluations. Each rater is accompanied by one of authors during evaluation task to keep a steady performance measure
- Human raters are allowed to listen to a specific utterance many times, but once a judgment is made, it cannot be changed. Each human rater has to evaluate 100 sets of utterances from different non-natives. In total, 500 sets of utterances from 5 non-natives students are evaluated by each human rater
- Evaluations are based on the understandability of each utterance. When understandable, it is accepted; otherwise, it is rejected. As a result, each human rater has the list of accepted-rejected utterances from 5 non-native students
- To make a final evaluation, the judgment of each utterance by the three human raters is evaluated by majority rule to lead a decision. When two human raters agree to accept a certain utterance while one human rater rejects it, voting is carried out to determine the result

Average of intra-rater reliability and inter-rater reliability for overall pronunciation is the same as those for segmental quality in knowledge-based method as the evaluation of overall pronunciation and segmental quality is carried out in parallel. Results of perceptual-based evaluation are of a total of 500 non-native utterances, 115 (23%) utterances are rejected with regards to overall pronunciation. These results will be used in the next step, alignment analysis with recognition results obtained by the system. Recognition results are aligned with perceptual-based evaluation in measuring Hit, False Alarm (FA), Miss, Rejection and Hit + Rejection rates:

(Hit + Miss + FA + Rejection) rates [%] = 100%

- Hit: The percentage of correctly recognized utterances. Both the system and the human raters accept the utterances
- False Alarm (FA): The percentage of incorrectly recognized utterance. The system accepts the utterances of which the human raters do not accept
- Miss: The percentage of incorrectly rejected utterances. The system rejects the utterances of which the human raters accept
- Rejection: The percentage of correctly rejected utterances. Both the system and the human raters reject the utterances
- Hit + Rejection: The percentage of correctly recognized and correctly rejected utterances. Both the system and the human raters accept and reject the utterances

RESULTS

Table 2 summarizes the alignment analysis of the baseline system and the baseline system adapted with three Speaker Adaptation (SA) techniques (MLLR, CMLLR and VTLN) respectively. As shown in the Hit + Rejection, the SA techniques provide gain about 4.2 points (68→72.2%) and about 2.3 points (15.4→17.7%) in the FA, corresponding to decrease in the Miss about 6.5 points (16.6→10.1%) and about 2.3 points (7.6→5.3%) in the Rejection when the system evaluated on non-native students. A positive improvement is also happened to native with reduction about 10.7 points (13.8→3.1%) in the Miss while keeping absolute rate in the Hit + Rejection. From the results, it can be seen that the SA techniques improve the recognition performance of native and non-native utterances. In other words, the performances of the baseline systems adapted with SA techniques are satisfactory.

Table 3 summarizes the alignment analysis of the baseline system adapted with Acoustic Model Adaptation (AMA) and the baseline system adapted with the combination of AMA and three SA techniques (MLLR, CMLLR and VTLN) respectively. It is shown that for the baseline system adapted with AMA, the Hit + Rejection increases 2.0 points (68→70%) over the baseline system when evaluated on non-native students. For native, the Hit + Rejection also increases about 2.9 points (86.2→89.1%). For the baseline system adapted with the combination of AMA and SA techniques (MLLR, CMLLR and VTLN), there is an improvement over the baseline system in the Hit + Rejection about 6.2 points (68→74.2%) and 3.6 points (15.4→19%) in the FA, corresponding to decrease in the Miss about 9.8

points (16.6→6.8%) and about 3.6 points (7.6→4%) in the Rejection when the baseline system evaluated on non-native students. For native, a positive improvement is performed with a gain about 12.7 points (86.2→98.9%) in the Hit and a reduction about 12.7 points (13.8→1.1%) in the Miss.

The acoustic model for the systems is originally built with a low rejection in order to give more encouragement for non-native students. However, this approach results in a relatively large proportion in false rejection (Miss) and False Acceptance (FA). Some experiments (the baseline system adapted with SA, the baseline system adapted with AMA and the baseline system adapted with the combination of AMA and SA) conducted for 500 non-native utterances yielded quite fair correct acceptance rates, Hit (66.9, 64 and 70.2% respectively) for very beginner level students. These results imply that more than half of the non-native utterances are correctly accepted. Moreover, the overall accuracy, that is, the percentage of correct acceptance and correct rejection (Hit + Rejection) is slightly higher (72.2, 70 and 74.2% respectively).

DISCUSSION

Perceptual-based evaluation of human raters is used as a standard against results of the system. Evaluation is based on the same test set and the results obtained are as follows:

- Non-native: Hit = 77%, Rejection =23% and Hit + Rejection = 100%
- Native: Hit = 100%, Rejection = 0 and Hit + Rejection = 100%

In general, the native and non-native results show that the baseline system adapted with SA techniques, the baseline system adapted with AMA and the baseline system adapted with the combination of AMA and SA techniques are comparable with each other relative to the baseline system in terms of Hit, FA, Miss, Rejection and Hit + Rejection. This can be explained by the fact that when the baseline system adapted with AMA, which already covered variants of pronunciation is combined with SA techniques, mismatch pronunciations between native and non-native utterances have been masked by model parameters. As an individual system, the baseline system adapted with the combination of AMA and MLLR slightly outperforms the other systems, as shown in bold in Table 3. Comparison between the baseline system adapted with the combination of AMA and MLLR and the perceptual-based evaluation for overall performance (the Hit + Rejection) results in 25.1 points difference (this number comprises FA and Miss) for non-native students and no difference performance for native.

A slight gain in Hit and FA with the corresponding reduction of Miss and Rejection can be explained by the fact that a speech recognition system will have FA and Miss fluctuated and overlapped in between Hit and Rejection. The issue of an acceptable level of FA depends largely on the application of the system. As the system is trained on a sharing data between native and non-native utterances, the native utterances can be used to define as the acceptance criteria, such that utterances from non-native subjects exceeding the acceptance criteria are accepted while utterances not exceeding the criteria are not accepted. However, in practice, the utterances from native and non-native subjects will overlap to each other on a certain degree, which means that the choice for a given criteria results in a combination of the Hit and the Rejection. In this case, perceptual-based evaluation is a goal standard in determining the validity of the system evaluation which can be represented in equation as follows:

- Hit of the system = Acceptance of the perceptual-based evaluation
- Rejection of the system = No Acceptance of the perceptual-based evaluation
- Miss of the system = 0 and FA of the system = 0

An accurate procedure of recovering the Miss to gain the Hit and recovering the FA to the Rejection still needs to be experimentally set up and investigated in more detail.

CONCLUSION

This study presents work on the proposed acoustic model adaptation for Indonesian language Utterance Training System (UTS) based on non-native utterances. The study achieved two objectives: (1) to provide the list of typical mispronounced phones together with their pronunciation variants made by non-native subjects in general that can be used as a corrective feedback to improve UTS performance and (2) to propose the acoustic model adaptation based on objective no. (1) and to use it in combination with speaker adaptation techniques. The proposed adaptation demonstrates its potential by showing a positive improvement on correct acceptance and correct rejection rate (Hit + Rejection) when it is evaluated on native and non-native utterances. The performance of the proposed acoustic model adaptation depends strongly on the effectiveness of state-clustering procedure to recover only in-

vocabulary words. In a future study, a confidence measures to be discriminate between in-vocabulary and out-vocabulary words will be investigated. It is also found that alignment analysis between recognition results of the system and perceptual-based evaluation of human raters has a potential to provide significantly confidence assessment for both native and non-native utterances.

REFERENCES

Al-Haddad, S.A.R., S.A. Samad, A. Hussain, K.A. Ishak and A.O.A. Noor, 2009. Robust speech recognition using fusion techniques and adaptive filtering. Am. J. Applied Sci., 6: 290-295. DOI: 10.3844/ajas.2009.290.295

Alia, M.A.K. and T. Al Mograbi, 2007. Investigation of an acoustic temperature transducer and its application for heater temperature measurement. Am. J. Applied Sci., 4: 294-299. DOI: 10.3844/ajassp.2007.294.299

Franco, H., K. Neumeyer, Y. Kim and O. Ronen, 1997. Automatic pronunciation scoring for language instruction. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 21-24, IEEE Xplore Press, Munchen, Germany, pp: 1471-1474. DOI: 10.1109/ICASSP.1997.596227

Gales, M. and S. Young, 2008. The application of hidden Markov models in speech recognition. Found. Trends Sign. Process., 1: 195-304. DOI: 10.1561/2000000004

Giuliani, D., M. Gerosa and F. Brugnara, 2006. Improved automatic speech recognition through speaker normalization. Comput. Speech Lang., 20: 107-123. DOI: 10.1016/j.csl.2005.05.002

Goronzy, S., S. Rapp and R. Kompe, 2004. Generating non-native pronunciation variants for lexicon adaptation. Speech Commun., 42: 109-123. DOI: 10.1016/j.specom.2003.09.003

Haraty, R.A. and O. El Ariss, 2007. CASRA+: A colloquial Arabic speech recognition application. Am. J. Applied Sci., 4: 23-32. DOI: 10.3844/ajassp.2007.23.32

Hariharan, R. and O. Viikki, 2002. An integrated study of speaker normalization and HMM adaptation for noise robust speaker-independent speech recognition. Speech Commun., 37: 349-361. DOI: 10.1016/S0167-6393(01)00039-5

Legetter, C.J. and P.C. Woodland, 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Lang., 9: 171-185. DOI: 10.1006/csla.1995.0010

Moustroufas, N. and V. Digalakis, 2006. Automatic pronunciation evaluation of foreign speakers using unknown text. Comput. Speech Lang., 21: 219-230. DOI: 10.1016/j.csl.2006.04.001

Neumeyer, L., H. Franco, M. Weintraub and P. Price, 1996. Automatic text-independent pronunciation scoring of foreign language student speech. Proceeding of the 4th International Conference on Spoken Language, Oct. 3-6, IEEE Xplore Press, Philadelphia, PA., USA., pp: 1457-1460. DOI: 10.1109/ICSLP.1996.607890

Oh, Y.R., J.S. Yoon and H.K. Kim, 2007. Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. Speech Commun., 49: 59-70. DOI: 10.1016/j.specom.2006.10.006

Shen, W. and D. Reynolds, 2008. Improved GMM-based language recognition using constrained MLLR transform. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 31-Apr. 4, IEEE Xplore Press, Las Vegas, NV., pp: 4149-4152. DOI: 10.1109/ICASSP.2008.4518568

Sundermann, D., H. Ney and H. Hoge, 2003. VTLN-based cross-language voice conversion. Proceeding of the IEEE Workshop on Automatic Speech Recognition and Understanding, Nov. 30- Dec. 3, IEEE Xplore Press, USA., pp: 676-681. DOI: 10.1109/ASRU.2003.1318521

Tan, T.S. and S. Hussain, 2009. Corpus design for Malay corpus-based speech synthesis system. Am. J. Applied Sci., 6: 696-702. DOI: 10.3844/ajassp.2009.696.702

Wester, M., 2003. Pronunciation modeling for ASR-knowledge-based and data-derived methods. J. Comput. Speech Lang., 17: 69-85. DOI: 10.1016/S0885-2308(02)00030-X