An Application of Modified Confusion Network for Improving Mispronunciation Detection in Computer-aided Mandarin Pronunciation Training

Jun Qi *, Ruiying Wei, Runsheng Liu
Department of Electronic Engineering, Tsinghua University, Beijing
qij08@mails.tsinghua.edu.cn

***Abstract*— In this paper, we propose an application of confusion network for Mandarin mispronunciation detection. Compared to former published works, which are proven to work effectively and robustly in detecting mispronunciation in word level and only successfully detect mispronunciation in sentence level in strictly small constrained search space, our modified confusion network based Computer-aided Pronunciation Training (CAPT) system is designded for exploring mispronunciation detections in sentence level with less constrained search space. Our benchmark test based on this improved CAPT system shows that in sentence level the mispronunciation detecting precision rate is up to 93% for substituted case and 98% for deleted and inserted cases, while the Recall rates for all three cases are above 94%.**

I. INTRODUCTION

The CALL (Computer-Assisted Language Learning) systems for improving language learners' Mandarin have gained wide popularity within the community during the recent fifteen years. One particular application of CALL system, named "Computer Aided Pronunciation Training" (CAPT), aims at supporting productive training by asking learners to read accordingly to a given prompt, pointing out pronunciation errors and bringing forth suggestions for improvements. For example, a learner may mispronounce the Mandarin word "wo3" as /m o3/ (mo3) and the CAPT system should be able to locate the mispronunciation error position and respond to the learner the error.

In general, Mandarin mispronunciation errors in sentence level can be classified into three sorts: (1) substituted mispronunciation error, e.g. "wo3"-"mo3", (2) deleted mispronunciation error, e.g. "wo3 hen3 hao3"-"wo3 hao3" and (3) inserted mispronunciation error: "ni3 hao3"-"ni3 bu2 hao3".

In the implementation of a CAPT system, the widely applied approaches can be classified into two categories: (1) use of confidence measures based on ASR, e.g. GOP [2] and Scaling Posterior Probability [3]; and (2) classification using other acoustic-phonetic features, e.g. LDA on formants and durations [4].

In our study, we focus on CAPT system improvement based on category (1). The most recent development of this sort of methods come from published work [1], where discriminative acoustic model has been used to produce one best sentence MAP hypothesis to detect mispronunciation errors through alignment between this recognized transcription and given prompt. However, the application of this method is strictly limited in sentence level. Our test shows that although it can successfully detect mispronunciation errors in word level with high accuracy, but it only generalizes to sentence level for mispronunciation detection in a strictly constrained search space.

This paper aims to address the bottleneck of problems of mispronunciation detection in sentence level. We propose an improved system with the technique of confusion network, which consists of sequential confusion sets with mutually exclusive hypothesis candidates. Being different from conventional processing of alignment between only one best sentence MAP hypothesis and given prompt, every given word in prompt should be aligned with candidates in one confusion set as shown in Figure 4,5,6,7 illustrated in section 2. In the process, the given word in prompt either corresponds to a confusion set with competing candidates or points to NULL which represents deleted mispronunciation. Besides, before the step of transcription alignment, the original generated confusion network needs to be modified to deal with "noisy hypothesis" whose definition will be explained in section 2.

The paper is organized as follows: the second section originally put forth our new CAPT system. The corpus preparation and corresponding test setup and experiment are proposed in the third section. The fourth part will give the final conclusion.

II. SYSTEM DESCRIPTION

A. *Our CAPT System Structure*

The basic structure of our system is shown in Fig.1. Compared with conventional baseline system, the component of Confusion Network Generation, which is drawn in dash line, is added to change lattice graph output into a compact representation named "confusion network" for the following transcriptions alignment between hypothesis confusion sets and given prompt.

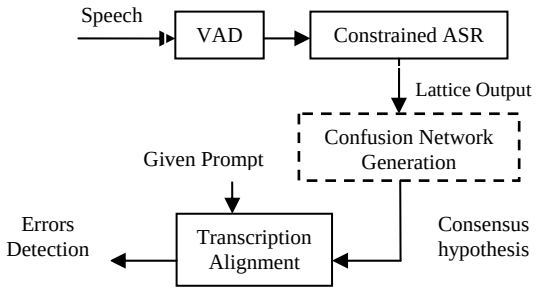

Fig. 1 Structure Diagram of our CAPT System

B. Components Description

(1) Constrained ASR

Based on the concepts from published work [6], the constrained ASR means that speech recognition is conducted with constrained language model (LM). For each frame in a segment corresponding to the phone q_i, we compute the frame-based log-posterior probability $P(q_i|\mathbf{o}_t)$ of the phone q_t as following:

$$\log(P(q_i \mid \mathbf{o}_t)) = \log(\frac{p(\mathbf{o}_t \mid q_i)p(q_i)}{\sum_{j=1}^{M} p(\mathbf{o}_t \mid q_j)p(q_j)}) \quad (1)$$

where $\mathbf{o}_t$ is the speech observation of phonetic segment. With constrained LM, the given word searching space is deliberately constrained to ensure sufficient high recognition rate for the following processing. Besides, ASR output can be stored in the format of lattice which could be useful for the following confusion network generation.

(2) Confusion Network Generation

In this section, a background introduction to confusion network is given. For a more in depth description and its detailed generation process, the reader can refer to [4].

The concept of confusion network is proposed by L.Mangu as early as 1998. Initially, the motivation of proposal of confusion network is to minimize word error rate (WER) by addressing the mismatch between the standard MAP paradigm which is sentence-based and the standard evaluation metric which is word-based. The basic idea of generation of confusion network is to extract high posterior probability word hypotheses from word lattices and find a complete alignment of all words in the lattice, identifying mutually supportive and competing word hypotheses. The generating steps of confusion network can be depicted in Fig.2.

Since the technique changes the standard problem formulation of searching among a large set of sentence hypotheses to a local search in a small set of word candidates, it indeed provides a more perspicuous representation of sequential sets of mutually exclusive word hypotheses, which provides significant hypothesis information for the following transcription alignment.

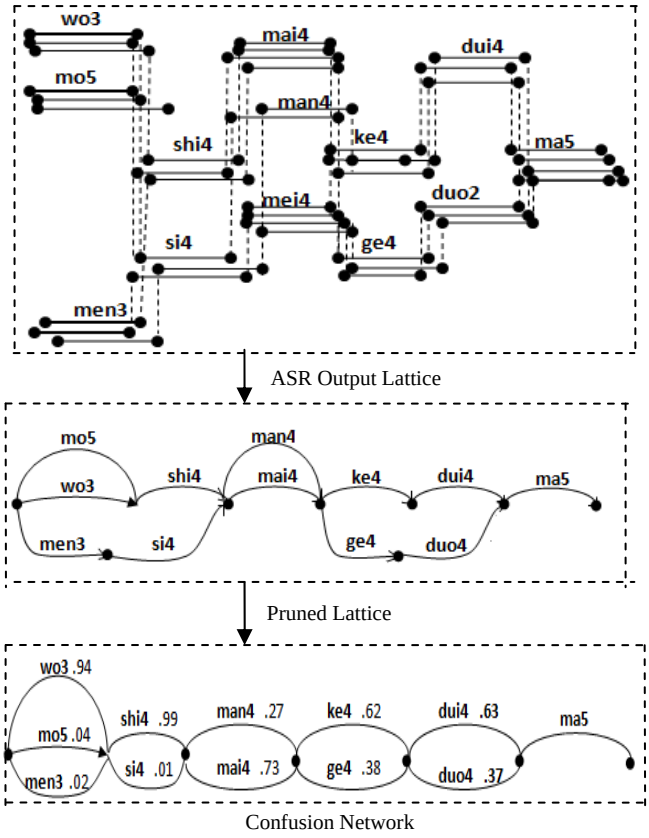

Fig 2. Generation of Confusion Network

(3) Transcription Alignment

The final step in our CAPT system is transcription alignment, which should be the key to final detection consequence. Conventionally, the detection depends on the alignment between one best sentence MAP hypothesis and given prompt. However, this approach to sentence-level mispronunciation detection is strictly limited in a very small searching space and this method tends to be no use when there are more searching candidates involved. One example of substituted mispronunciation detection is shown in Fig.3, where the first line is the given prompt and the second and third line are the best MAP sentence hypotheses when the sizes of search space are separately 25 and 100 Mandarin monosyllable words.

Transcription:	wo3 shi4 mai4 ke4
One best hypothesis:	
(25 words)	wo3 shi4 mai4 ke4
(100 words)	mo5 shi4 hai4 te4

Fig.3 Transcription Matching using 1-best MAP hypothesis

In fact, the speaker's utterance is perfectly according to the give prompt. But the one best sentence MAP hypothesis cannot be effectively used for transcription alignment as the low recognition rate is concerned. The reason is due to the fact that the Viterbi beam searching algorithm, which is widely used in speech recognition, is to search for the best states path with the maximum likelihood as the whole sentence MAP hypothesis, but it cannot guarantee every local state path is

optimal. It means that some word hypotheses in an optimal sentence hypothesis are suboptimal. One approach to dealing with the problem is to output N-best sentence hypotheses. However, it is not unusual to find that N best consequences always target at one word choice, provided that this word has many competing candidates with similar likelihoods. Besides, the format of N-best hypotheses consequence is not a compact representation for information post-processing.

Since confusion network can provide more useful recognition information including mutually exclusive word hypotheses and likelihood assigned for each hypothesis candidate in each hypothesis set, the technique of confusion network can be assumed to be useful for addressing the problem.

Based on the application of confusion network, the process of transcription alignment can be modified as Fig.4. Being different from one best sentence hypothesis based hypothesis, which only provides one choice, the confusion network based one provides alternative competing hypothesis candidates with likelihoods for transcription alignment. For simplicity, we can choose the candidate with the highest likelihood for alignment with given word in prompt. Based on the method, “wo3” and “mai4” are successfully matched with prompt, but “ke4” is still misrecognized as “te4”.

To handle this problem, an improved method is to not only consider the candidate in one confusion set with maximum likelihood, but also involve other alternatives with relatively high likelihood in the set. In our experiment, we choose the top three candidates in one set for alignment. In doing so, the word “ke4” is correctly matched with given word in prompt. A further processing is to inform the learner that his word pronunciation is highly correlated with word “te4” and please pay more attention to this word pronunciation.

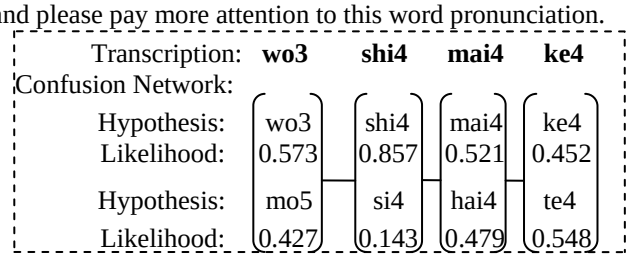

Fig.4 Transcription Alignment with Confusion Network for substituted case

The example mentioned above belongs to the substituted mispronunciation case in sentence-level. A more complex problem belongs to the deleted and inserted mispronunciation categories. But firstly a modified confusion network is needed to overcome “noisy hypothesis” and then point out existed mispronunciation errors.

“Noisy hypothesis” can be defined as the background stationary noise which is mixed with speech and is misrecognized as one or more word hypotheses with likelihood, especially when the word searching space is extended. In the case of “noisy hypothesis”, confusion network is of great significance as de-noising filter. As shown in Fig.5, since the noise is assumed to be stationary, “noisy hypothesis” has to compete with “delete” in all cases.

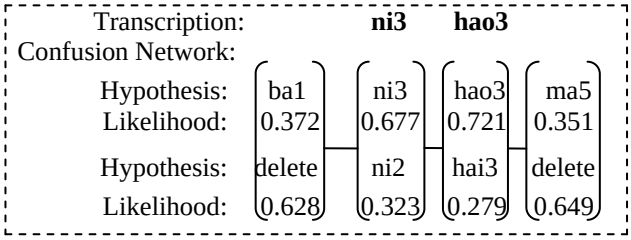

Fig.5 “noisy hypothesis” cases

For processing “Noisy hypothesis”, confusion network is needed to make some modification to overcome it. One simple approach is to look for all the confusion sets with key word “delete” and removed those sets. This method is reasonable for one main reason: the background noise in our test environment is assumed to be stationary and thus there is no more than one misrecognized candidate in searching space, so the likelihood of “noisy hypothesis” is less than 1. After processing the “noisy hypothesis”, the detection of inserted and deleted mispronunciation starts to work. As shown in Fig.6, the inserted mispronunciation error “bu2”, which has higher likelihood than another inserted candidate word “duo1”, is detected as inserted mispronunciation.

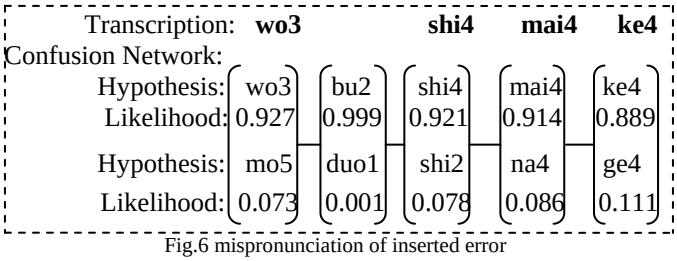

Fig.6 mispronunciation of inserted error

Another instance as shown in Fig.7 belongs to deleted mispronunciation type, where the given word in prompt cannot be aligned with any confusion set. As shown in Fig.7, “ke4” is detected to be not pronounced.

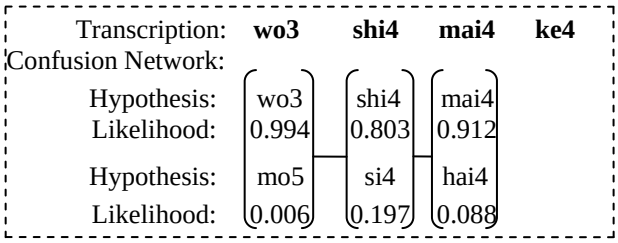

Fig.7 mispronunciation of deleted error

III. EXPERIMENT

A. Corpus Preparation

Our research is based on the Tsinghua-Mandarin corpus, which contains recordings of 24 non-native Chinese-speaking learners of Mandarin. Each leaner need to read completely same 186 sentences based on given prompts. Those given prompts include both correct sentences, which are seen as the prompt, and sentences involving all types of pronunciation errors, which are artificially made. So in the test, each learner should read all the given correct and incorrect sentences based

on reading text. Totally, there are 4464 sentences which are needed to be detected. The statistics of all types of mispronunciation errors of the given 186 sentences for every speaker are listed in Table I, where 47 sentences of the total are shown as reference and others are artificially generated error sentences.

Total test sentences	186
Sentences as prompts	47
substituted errors	92
Deleted errors	36
Inserted errors	43

Table I: statistics of transcription for one speaker

B. Experiment setup

In the front-end, a 45-dimensional feature vector is extracted from the test speech, including 14-dimensional MFCCs with normalized log-energy and their first and second order differentials. Since Mandarin is a language with 5 tones, a 3-dimensional tone feature vector is appended to the spectral features, resulting in a final feature vector of 49-dimension.

The acoustic model is trained by using HTK3.4, and is based on cross-word triphones modeled by 3-state left-to-right HMMs. A decision-tree based state tying is applied resulting in a total of 2400 triphone states. The state output densities are 16-component Gaussian mixture models with diagonal covariances. The training dataset is based on 863 and Microsoft Standard Mandarin speech corpus.

Considering the generalization of our proposed approach, uni-gram (word-loop) language model (LM) is chosen for our experiment and only word coverage of LM is varied.

C. Experiment results

(1) Benchmark test

For the benchmark test, we alleviate word search space only covering all the used Mandarin monosyllable words in test Mandarin corpus, the total number of which is 112.

Comparing with the results of detection based on 1-best MAP sentence hypothesis alignment, our CAPT results in a significant improvement to mispronunciation detection. As shown in Table II, the detecting precision rates of all three types are above 93% while the recall rate is as high as 94%. The detection rates based on confusion network in this LM setup can be considered as benchmark for following experiments.

	Substituted	Inserted	Deleted
Total number	2208	1032	864
System based on Confusion Network			
Precision rate	93.8%	97.6%	98.6%
Recall rate	94.5%	99.4%	99.19%
system with one best sentence hypothesis			
Precision rate	48.6%	50.5%	56.3%
Recall rate	41.2%	37.6%	72.2%

Table II: Benchmark result of our system

(2) Further exploring tests

A further explore of our CAPT system lies in the point that how to extend word searching space as large as possible, while maintaining high mispronunciation detection rate. In LM setup, we gradually increment searching space by one hundred syllables which are randomly selected from totally 1200 standard Mandarin word syllables. The average results of many round of tests based on 200, 300, 400 size are separately listed in Table III.

	substituted	Inserted	Deleted
Total number	2208	1032	864
200 monosyllabic words			
Precision rate	89.6%	93.6%	94.6%
Recall rate	91.23%	95.38%	97.26%
300 monosyllabic words			
Precision rate	83.4%	87.7%	90.3%
Recall rate	84.4%	92.7%	94.2%
400 monosyllabic words			
Precision rate	72.8%	78.1%	81.9%
Recall rate	72.6%	84.4%	86.1%

Table III: test results based on LM of 200, 300 and 400 monosyllabic words

As shown in above Table III, the correct detection rates drop rapidly when the size of search is set to be 400 words. It means that in the case of unigram LM, the allowable searching space of our confusion network based CAPT should be within 300 words. Otherwise, the detection results cannot be guaranteed to have high confidence score due to low detection rate. Even so, our new CAPT has achieved significant improvement, since it greatly extends word search space and thus put less constraint on LM.

IV. CONCLUSION

The Mandarin mispronunciation detection to all types of errors in sentence level is an important issue in CAPT. Since there are totally no more than 1200 Mandarin monosyllable words, the issue seems to be more promising. In this paper, we propose a CAPT system based on confusion network and show that it can significantly improve the performance of mispronunciation detection for Mandarin monosyllable words. Compared with the conventional one best MAP sentence hypothesis for transcription alignment, our confusion network based one can tremendously alleviate the constraint of language model while the size of search space is greatly extended. In practice, our test shows that our CAPT system is convinced to be an important step forward for real application, specifically for Mandarin mispronunciation detection.

REFERENCE

[1] Qian, X.J., Frank S., Helen M., “Discriminatively Trained Acoustic Model for Improving Mispronunciation Detection and Diagnosis in Computer Aided Pronunciation Training (CAPT)” in INTERSPEECH, 2010.

[2] Meng, H, Lo, Y.Y., Wang, L. and Lau, W.Y., “Deriving satient learners’ mispronunciation from cross-language phonological comparisons”, in ASRU, 437-442, 2007.

[3] Harrison, A.M., Lo, W.K., Qian, X.J. and Meng, H., “Implementation of an Extented Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training”, in SIG-SLATE, 2009.

[4] Lidia Mangu, Eric Brill, Andreas Stolcke, “Finding Consensus among Words: Lattice-based Word Error Minimization”, in Eurospeech, 1999.

[5] L.R.Bahl, F.Jelinek, and R.L.Mercer, “A maximum likelihood approach to continuous speech recognition”, In IEEE Transactions on Pattern Analysis and Machine Intelligence, volume PAMI-5, pages 179-90, March 1983.

[6] Shi Q., Zhang S.L., Stephen M.Chu, Xiao, J., Ou Z.J.,”Spoken English Assessment System for Non-Native Speaker Using Acoustic and Prosodic Features”, In INTERSPEECH, 2010.

[7] Harrison, A.M., Lau, W.Y., Meng, H. and Wang,L., “Improving mispronunciation detection and diagnosis of learners’ speech with context-sensitive phonological rules based on language transfer”, in INTERSPEECH, 2008.

[8] L. Wang, X.Feng and H.Meng, “Automatic Generation and Pruning of Phonetic Mispronunciation to Support Computer-Aided Pronunciation Training.”, In INSPEECH, 2008.

[9] L. Chen, K. Zechner, and X. Xi, “Improved Pronunciation Features for Construct-driven Assessment of Non-native Spontaneous Speech”, in proceeding of NAACL, Boulder, 2009.

[10] Yin, Shou-Chun, Richard Rose, Oscar Saz, and Eduardo Leida. “A Study of Pronunciation Verification in a Speech Therapy Application”, ICASSP-Robust Speech Recognition III. Taiwan, 2009.