ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

FEATURE MASKING IN AN EMBEDDED MANDARIN SPEECH RECOGNITION SYSTEM

Yuezhong Tang, Xia Wang, Yang Cao, Feng Ding
Nokia Research Center, Beijing

ABSTRACT

In this paper, we explored a feature component masking scheme for an embedded tonal language recognition systems, in order to reduce the computational complexity with least degradation of recognition accuracy. We made a lot of experiments on a Mandarin isolated word recognition task with a tone-confusable vocabulary. Taking consideration of both clean and noisy conditions, we were able to find a masking scheme that filtered out 31 of 54 components and still outperformed the baseline with 54 components in feature set, with dramatically less computational and memory complexity. The results showed that feature masking was a promising approach for complexity reduction in embedded tonal language recognition systems. The results also verified the effectiveness of higher order cepstral coefficients for tonal language recognition because most of them were preserved during the feature masking experiments.

1. INTRODUCTION

Automatic Speech Recognition (ASR) is resource demanding. With the development of hardware, ASR system can be implemented in embedded devices such as PDAs or cell phones. However in these platforms, the computational power and memory resources are still very limited which restrains speech technology going to mobile. It is therefore necessary to explore techniques to reduce complexity of ASR systems for embedded platforms.

Such techniques fall into three main categories: feature component masking, variable-rate partial likelihood update and density pruning [1]. In this paper, we focus on feature masking method, which aims at the reduction of feature dimension.

Our target language is a tonal syllabic language, Mandarin Chinese, which requires more information in ASR than Western languages. On the one hand, integration of tonal information is very important for Mandarin speech recognition; on the other hand, it will take additional computational and memory cost.

There are two typical approaches of using tone information in ASR systems. One is to take the pitch value into consideration in feature extraction and treat it as part of the feature vector coefficients [2, 3, 4, 5]. Another approach is to build a separate tone classifier[6]. Both methods rely on accurate estimation of pitch value from the speech data, which is still a challenging task, especially in noisy environments. We have found that Higher Order Cepstral (HOC) [2, 3] features can significantly improve the tone recognition accuracy. Our experimental results showed that HOC features could improve tonal Mandarin Chinese ASR system, but did not help in Western language ASR system.

At the same time, the memory consumption and computational complexity increased unavoidably. To take the best use of HOCs and keep the complexity at a level comparable to traditional 39 MFCC components, we explored the feature masking approach for Mandarin speech recognition on a isolated tone-confusable vocabulary.

The remainder of the paper is organized as follows. Section 2 describes our speaker-independent ASR system with tone recognition. Section 3 continues with the proposed feature masking method. Section 4 highlights the experimental results and finally, conclusions and outlook are given.

2. SPEAKER-INDEPENDENT ISOLATED WORD RECOGNITION WITH TONE RECOGNITION

We proposed the Tiger system in [2], in which tone information was carried out by main vowel phonemes. Tiger was a mono-phone system because in embedded platforms, context-dependent modeling scheme is too expensive. HOCs were used for inexplicit tone information modeling based on our previous research results of its excellent performance (10 percentage units higher accuracy). In the feature masking research, we took Tiger with 54 feature components as a baseline system.

The training corpus consists of speech data from 50 male and 50 female speakers. It was recorded in a quiet office, 8 kHz sampling rate, 16-bit coding, consisting of isolated words and sentences. The total number of training utterances was more than 100,000.

The Mandarin in-house name database "Pitch99" was used for testing. Pitch99 is designed for the verification of pitch estimator and tone classifier. It consists of 64 pairs of easily confusable Chinese names. The two names within a pair are identical if the tone is ignored. Here are two examples, "fu4 zuo4 yi4" vs. "fu2 zuo4 yi4", and "bao1 yun2" vs. "bao4 yun2". Ten male and ten female speakers took part in recordings in a quiet office environment. Those data formed the original clean database. Noisy database is obtained by adding car, cafe or music noise to clean utterance. The signal-to-noise ratio of the noisy database is uniformly distributed between 5 dB and 20 dB.

The phoneme set is defined based on the SAMPA-C definition[7], and some language-specific modifications have been made to

increase the modeling accuracy. Each phoneme is modeled with a 3-state HMM, in which each state is modeled with 8 mixtures. Main vowel phonemes have 4 variants representing 4 tones of the same phoneme.

The feature vector in recognizer composes of 17 cepstral coefficients obtained from 30 mel-bands, energy and their first-order and second-order time derivatives, ending up with 54 components, in which 39 are standard Mel-Frequency Cepstral Coefficients (MFCC) and the rest higher order coefficients are called HOCs.

Among the 54 components, extensive experiments were made to find out the optimum subset of the feature vector components that balances the number of components and the recognition accuracy. Based on the experimental results, we could provide flexible footprint system according to the requirements of complexity and accuracy. Interestingly but not surprisingly, we found only few HOCs were filtered out during all the experiments (e.g. when 31 components were masked, only 4 of them came from higher-order coefficients) and this result, from another angle, proved the effectiveness of HOCs for speech recognition with tones.

3. EXPERIMENT DESIGN

It is a fact that different feature vector components contribute differently to the overall recognition performance. Masking some of the less important components can save computation and memory consumption without much degradation of recognition accuracy.

Compared with other methods, feature masking is simple and straightforward. The key here is to define a process and criteria for filtering or masking, namely how to find the less important features. Let's consider the accuracy with n components masked ($0 \le n < 54$),

1. n=0: no component is masked. The recognition accuracy is defined as baseline performance with all feature components involved in probability calculation in the decoding phase.
2. n=1: only 1 component is masked. The recognition accuracy is defined as the best result among the 54 implementations with one dimension was ignored each time.
3. $1<n<54$: n components are masked. There are C_{54}^{n} kinds of masking possibilities with different combinations of n dimensions filtered out in the decoding phase. The number is huge for a certain range of n, so we decided not to target at the global maximum. Instead of making C_{54}^{n} implementations, we did in an iteratively way to take the best 18 implementations of masking length = n-1, and filter out one more dimension. The best recognition result of the $18 \times C_{54-n+1}^{1}$ or less (if there are identical schemes) masking implementations is defined as the accuracy of masking length of n.

We started from n=1 to see the individual importance of each dimension, then expand to n=2, n=3, ... Because the components are not independent, there is no straightforward relationship between the sub feature set of M ($0<M<53$) and M+1 dimensions. The goal is to mask as many as possible feature dimensions while keeping the recognition performance close to the baseline.

As you may noticed already, the experimental scheme described above could not lead us to the global optimal feature set and we found it too. Personal judgment and analysis of the past experiments were also used for new experiment scheme design. With manual check, we were able to find much better results with much longer masking length as shown in Table 5, 6 and 7.

During the experiments, we treated HOCs in the same way of the left 39 dimensions of MFCCs. According to our past experience, HOCs should play important role in tone classification and the results verified this by showing that in those best masking schemes, most of HOC features were preserved. Even at an extreme case, when 31 dimensions were filtered out, among which there were only 4 HOC features.

4. EXPERIMENTS RESULTS

In this section, we will present various experimental results, with feature masking applied on clean data, noisy data, with / without speaker adaptation to show the effectiveness of the feature masking scheme and verify the validness of higher order cepstral coefficients.

Experiments in 4.1 – 4.4 were carried out using the 3-step process described in section 3 and the results might not be optimal globally. However, those results are enough to show the effectiveness of feature masking schemes. Experiments in 4.5 took the advantages of our empirical knowledge by manual check and analysis of the experimental results; therefore they achieved much better results than those in 4.1-4.4.

4.1 Clean without Adaptation

Table 1 below shows part of results we obtained on clean database without any speaker adaptation. Only those better than baseline are listed, in descent order of recognition accuracy.

Table 1 Results of clean / no adaptation

Masking length	Accuracy (%)
0: Baseline	87.88
9	88.11
11	88.11
8	88.05
5	88.04
6	88.03
12	87.99
7	87.98
13	87.96
10	87.92
14	87.91

It is obvious that 1) with less features, we could obtain even better results, i.e. number of features is not the key, but quality of features is; 2) 14 features could be left out at most while the performance is still better than the baseline.

One fact not visible from Table 1 is that no HOC feature was masked in the above experiments.

4.2 Clean with Speaker Adaptation

When on-line Bayesian speaker adaptation was adopted on the clean database, we got even better results. Again, fewer features could obtain better results and HOCs were preserved during the experiments. However, the best masking scheme with the same masking length n in Table 2 is not necessarily the same with the one in Table 1, and most probably, they are different.

TABLE 2 Results of Clean With Speaker Adaptation

Masking length	Accuracy (%)
0: Baseline	88.29
11	89.57
13	89.42
12	89.36
10	89.30
7	89.28
14	89.24
6	89.06
8	88.98
5	88.87

The result of this experiment is similar to experiment 1. Most of the results are better than the baseline. The length of masking features is not related with the performance. But the concrete masking scheme and performance rank are different from the first experiment.

4.3 Noisy without Adaptation

TABLE 3 Results of Noisy Without Speaker Adaptation

Masking length	Accuracy (%)
0: Baseline	82.28
5	82.12
4	82.04
6	81.39
7	81.18
8	81.04
10	80.93
9	80.88
12	80.68

Table 3 shows the best 8 feature masking results on the additive noisy database without speaker adaptation. Here we observed different characteristics of feature masking schemes under adverse conditions. On noisy database, without any exception, all masking schemes lead to degradation of recognition performance to some extent. Further more, it is roughly true that more features used, better results obtained. This means that some features are essential in noisy environment, but not so important or redundant in clean. In these results, HOC features were preserved also.

4.4 Noisy with Speaker Adaptation

Speaker adaptation obtained much more significant improvement of accuracy on noisy database, for both baseline, and masked feature set. And after adaptation, the characteristics of the masking scheme turned out to be more like the cases in clean environment and we were able to find different combinations that were better than the baseline after adaptation with 12 dimensions filtered out. As in the other experiments, HOCs remained to be important features here.

TABLE 4 Results of Noisy With Speaker Adaptation

Masking length	Accuracy (%)
0: Baseline	87.22
7	87.34
10	87.29
12	87.29
6	87.26
5	87.22
4	87.12

4.5 Comprehensive Experiments

Given a criteria of number of masking features, or preferred recognition accuracy, the four kinds of experiments above will lead to an optimal feature set. There is no perfect solution that suits for all conditions, so we need to find a balance between noisy and clean environment, with and without speaker adaptation. We assume that it is easy to set an option to switch on/off speaker adaptation in the user interface design speech recognition applications running on mobile terminals, therefore we only need to balance clean vs. noisy performance when speaker adaptation is "on" or "off".

On another hand, we tried to use our knowledge and experience to select good candidates in the experiments with combined clean/noisy environments and the results are listed below.

4.5.1 Without Adaptation

In these results, we treat clean and noisy equally. According to different requirements, we could assign different weights for recognition performance in clean and noisy environments.

Table 5 shows the best feature masking results when both clean and noisy database were taken into consideration and no speaker adaptation was adopted.

TABLE 5 Comprehensive Results Without Speaker Adaptation

Masking length	Clean (%)	Noisy (%)	Avg (%)
0: Baseline	87.88	82.28	85.08
6	87.96	82.10	85.03
7	87.94	81.95	84.95
5	87.88	82.00	84.94
8	88.02	81.80	84.91
9	87.80	81.77	84.79
10	86.99	81.22	84.11

Without adaptation, as we saw from Table 1 and Table 2, although some masking schemes may bring better results in clean, they

cannot compensate the loss in noisy environment. With an equal weight of noisy and clean results, we were not able to find a solution better than baseline. However, when 8 features were filtered out, i.e. the accuracy degradation was marginal and the masking scheme should serve as a good solution in embedded systems that suffer from limit resources. In noisy column, you may find some results are better than those in Table 3 with same masking length, showing the benefits of incorporating empirical knowledge.

4.5.2 With Speaker Adaptation

With the excellent performance of speaker adaptation, we found a lot of combinations that outperformed the baseline feature set. And among those masking schemes, most HOC features were preserved. One surprising result was that when there were as many as 31 features (57% of the whole feature set) masked, the masking scheme still outperformed the baseline, which was very promising and encouraging. Only 4 out of 31 masked features were from HOCs, in which 1 from 0th-order, 1 from 1st-order, 2 from 2nd-order derivatives. This result indicates that we could make good use of different masking schemes to generate much more flexible footprint ASR systems according to the resources available.

TABLE 6 Comprehensive Results With Speaker Adaptation

Masking length	Clean (%)	Noisy (%)	Avg (%)
0: Baseline	88.29	87.22	87.76
9	90.89	88.54	89.72
11	90.90	88.52	89.71
10	90.73	88.53	89.63
8	90.56	88.44	89.50
7	90.41	88.56	89.49
20	91.17	87.21	89.19
27	91.18	86.78	88.98
21	90.97	86.90	88.94
24	90.92	86.91	88.92
28	91.28	86.54	88.91
26	91.21	86.41	88.81
29	91.17	86.21	88.69
30	91.35	85.90	88.63
31	91.60	85.53	88.57

5. SUMMARY

From the results in Section 4, we can draw a conclusion that feature masking is an effective method to reduce the computational complexity with even better recognition results or marginal degradation. The promising results in Table 6 showed that we could reduce the complexity dramatically when needed. However, there is no universal perfect masking scheme for all usage situations (noisy, clean, adaptation, etc.). Even in the same condition, the reduction approach is nonlinear. For example, if A, B and C are three masking schemes, and the performance of A is better than B, the performance of A+C is not necessarily better than B+C. This non-linearity is related with the nonlinear nature of the features.

In our experiments, higher order features continue to play important roles in tone classification because few of them were filtered out in the feature masking experiments. Through these results, the effectiveness of HOCs for tonal language recognition was verified.

6. OUTLOOK

Our work so far shows that the feature masking is a promising approach to reduce computational complexity and memory consumption for embedded systems. However, because of the nonlinear nature of the feature components, our search algorithm was not globally optimized. The adjustment of masking scheme is somewhat subject and depending on the researcher's personal judgment.

From the results listed in Section 4, we're now pretty confidant that feature masking is a good way to go for complexity reduction in embedded systems. More systematic and objective experiments will be designed and carried out to find the globally optimal solutions for different environmental conditions. More investigation needs to be done also to find the relationship between different feature components.

In our recognition database, the vocabulary is composed of 64 pairs of tone-confusable names, which is a quite tough task. More realistic vocabulary with moderate tone-confusable entries shall be used in our future work.

It is also interesting to test feature masking on a real noisy database for noise robustness research rather than on an artificially created one.

REFERENCES

[1] I. Kiss, M. Vasilache, "Low complexity techniques for embedded ASR systems", ICSLP-2002, Denver, 2002

[2] X. Wang, J. Iso-Sipilä, "Low Complexity Mandarin Speaker-Independent Isolated Word Recognition", Proc. of ICSLP 2002, Denver, 2002

[3] X. Wang, Y. Dong, J.Iso-Sipilä, O. Viikki, "On integrating tonal information into Chinese speech recognition", ISCSLP, Beijing, 2000

[4] C.J. Chen ,et al. "New methods in continuous Mandarin speech recognition", Eurospeech Proceedings, 1997

[5] C.H. Huang, F. Siede "Pitch tracking and tone features for Mandarin speech recognition", ICASSP2000, vol.3, pp1523-1526, 2000.

[6] H.M. Wang, T.H. Ho, R.C. Yang, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data", IEEE Trans on Speech and Audio Processing, Vol 5, No 2, pp 195-200,1997

[7] SAMPA-C definition developed by Chinese Academy of Social Sciences, http://www.cass.net.cn/s18_yys/yuyin/sampac/sampac.htm