AUTOMATIC DERIVATION OF A PHONEME SET WITH TONE INFORMATION FOR CHINESE SPEECH RECOGNITION BASED ON MUTUAL INFORMATION CRITERION

Jin-Song Zhang, Xin-Hui Hu, and Satoshi Nakamura

ATR Spoken Language Communication Research Laboratories
2-2-2 Kansai Science City, Kyoto 619-0288 Japan
{jinsong.zhang, xinhui.hu, satoshi.nakamura}@atr.co.jp

ABSTRACT

An appropriate approach to model tone information is helpful for building Chinese large vocabulary continuous speech recognition system. We propose to derive an efficient phoneme set of tone-dependent sub-word units to build a recognition system, by iteratively merging a pair of tone-dependent units according to the principle of minimal loss of the mutual information. The mutual information is measured between the word tokens and their phoneme transcriptions in a training text corpus, based on the system lexical and language model. The approach has the capability to keep discriminative tonal (and phoneme) contrasts that are most helpful for disambiguating homophone words due to lack of tones, and merge those tonal (and phoneme) contrasts that are not important for word disambiguation for the recognition task. This enable a flexible selection of phoneme set according to a balance between the MI information amount and the number of phonemes. We applied the method to traditional phoneme set of Initial/Finals, and derived several phoneme sets with different number of units. Speech recognition experiments using the derived sets showed their effectiveness.

1. INTRODUCTION

Chinese is a tonal language, in which each syllable is associated with a kind of pitch tone. There are four basic tones and one neutral tone. The same syllables with different tones have different lexical meaning. It has been an interesting and important topic how to model the tone information to build a Chinese large vocabulary continuous speech recognition (LVCSR) system. Among a number of various kinds of approaches, the one using tone dependent sub-word units has the advantage of frame-synchronous consistency with the decoding strategy of the state-of-art LVCSR system, and has been widely adopted [1, 2, 4]. One common problem of these approaches is that the number of phoneme set of the LVCSR system will increase significantly after introducing tone dependencies. For ex., in the case of widely used traditional Chinese phoneme set of Initials/Finals(IFs), the number of non-tonal IFs is 59, and that of tone-dependent ones is more than 200. As context dependent tri-phone HMMs are usually used in LVCSR systems, their number will explode from tens of thousands to millions when tone-dependency is used, making it very challenging how to train the tri-phone HMMs robustly. Also, the complexity of the phoneme hypotheses lattice will increase significantly, making the decoding more computationally heavy.

The approaches to deal with the problem in the previous studies [1, 2, 4] are to hand-craft a small phoneme set containing tone-dependent phonemes, like tonemes [1], tonal main vowels [4], segmental tones [2] and etc.. Although they showed performance improvements in the recognition experiments, they still need to increase the phoneme set by several times due to a full expansion of non-tone units to tone dependent ones. However, a full expansion of tone dependencies may be unnecessary. On the one hand, speakers tend to reduce some tones from their lexical forms in daily speech [5] when the reductions do not obstacle speech communication. On the other hand, the lexical and language model (e.g., n-gram) information in an LVCSR system is usually very efficient to disambiguate most of homophone words due to a lack of tone information [6], as evidenced by the fact that an incorporation of several-times-big tonal phoneme set has led to only slight recognition improvements[1, 4].

By viewing the full expansion of tone dependencies as unnecessary, we propose that only those tone dependencies be incorporated that are necessary for disambiguating word confusions of an LVCSR system. Different from several previous studies on disambiguating word confusions which are based on the acoustic confusions of phonemes [7, 8, 9, 10], our method focuses on the disambiguation power from the lexical and language model. In other words, a tone dependency is not incorporated when the lexical and language model can disambiguate those homophone words resulting from the lack of that tone.

The real approach is realized as compacting the redundancy of an initial full-tone-dependent unit set, according to the principle of minimal loss of the mutual information. The mutual information is measured between the word tokens and their phoneme transcriptions in a training text corpus. A greedy search is adopted to merge two units at a time to minimize the corresponding mutual information loss. The final phoneme set can be flexibly chosen according to a balance between the number of units and the information quantities. Preliminary speech recognition experiments have been carried out to testify the effectiveness of the deduced phoneme sets.

2. CHINESE PHONOLOGY

A Chinese word is composed of one to several characters, and each character is pronounced as a monosyllable with a pitch tone. The totally phonetically differentiable tonal syllables are about 1,300, and the number of base syllables is about 410 when pitch tones are discarded. Traditional Chinese phonology divides the syllables into demi-syllabic units: 21 *Initials* and 37 *Finals*, plus four basic lexical tones(Tone 1-4) and a neutral tone (0) (Table 1).

Initials	b, p, m, f, d, t, n, l, z, c, s, zh, ch, sh, r, j, q, x, g, k, h, *null initial*
Finals	a, ao, ai, an, ang, o, ou, ong, e, ei, en, eng, er ia, iao, ie, i1, i2, i3, iu, in, ian, ing, iang, iong, u, ua, uo, ui, uai, un, uan, uang, v, ve, vn, van
Tone	0, 1, 2, 3, 4

Table 1. Pinyin symbols for Initials, Finals and lexical tones of Chinese syllables.

In the tone-dependent phoneme approach, all the Finals are expanded into tone dependent ones, like *a0, a1, a2, a3, a4* and etc.. Although not all the combinations of Finals and tones exist, the number of the tone dependent phoneme set is still more than 200. When isolate monosyllable words are considered, tone contrasts may play an important role in discriminating the words. For ex, the following words: *ma1*(mother), *ma2*(hemp), *m3*(horse) and *ma4*(scold), are only differentiated by the tones when in isolations. However, when in sentences, they will have very different context words. In other words, the lexical and language model (n-gram) has the power to disambiguate the four words even the tone information is ignored. Therefore, we regard that there are many redundancies in the original tone dependent IFs set for a recognition system, when given a lexical and language model.

3. THEORY FOR PHONEME SET OPTIMIZATION

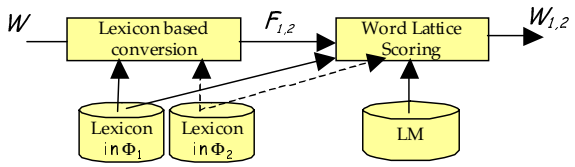

Fig. 1. Illustration of the problem formalization.

We formalize the phoneme set optimization problem into an information coding/decoding approach as illustrated in Figure 1, where W stands for word based text corpus, Φ_1 and Φ_2 for two different phoneme sets, $F_{1,2}$ for the different phoneme transcriptions of the W based on $\Phi_{1,2}$ lexicons respectively, $W_{1,2}$ for the decoded words from $F_{1,2}$ based on the same language model and the respective lexicons. When a coding method Φ_i is lossless, the decoded words W_i should satisfy $W = W_i$. However, when the coding is not uniquely decodeable, a better coding Φ^* should be the one

$$\Phi^* = \arg\max_i I(W, F_i) \text{ where } i = 1, 2$$

The mutual information $I(W, F_i)$ can be calculated as

$$\begin{aligned} I(W, F_i) &= H(W) - H(W|F_i) && (1)\\ &= \log P(W|F_i) - \log P(W) && (2)\\ &= \log \frac{P(F_i|W)}{\sum_{\text{all } j} P(F_i|W_j)P(W_j)} && (3) \end{aligned}$$

$P(W)$ and $P(F_i|W)$ represent two main components in the current speech recognition system: i.e., language modeling and probabilistic pronunciation variation modeling.

4. MINIMUM MI LOSS BASED PHONEME SET REDUCTION

We have designed a greedy approach to compact the redundancies of an initial phoneme set by iteratively merging one pair of phonemes whose merge leads to the least loss of MI. Figure 2 illustrated the flow chart of the method.

- Initialization condition: the following resources are prepared.
  - Initial phoneme set Φ_0: it contains the full tone-dependent sub-word units.
  - Lexicon: the one for speech recognition task and is represented in the initial phoneme set.
  - Text corpus: the one standing for the speech recognition task.
  - Language model: the one of the speech recognition system.
- Optimization procedure:
  1. MI calculations: for each possible merge of two phonemes $\Psi_i : A + B \rightarrow A$, the reduced MI $\Delta MI(\Psi_i)$ is calculated using the assumingly merged lexicon.
  2. Merge decision: among all the possible merges, the one Ψ^* that has the smallest reduced MI is selected as the effective merge of this iteration.

     $$\Psi^* = \arg\min_{\text{all } i} \Delta MI(\Psi_i)$$

  3. Renew the lexicon and phoneme set based on the effective merge Ψ^*.
  4. Check if the stop criterion is satisfied or not. If no, go to step 1 and do 1-3 once again. If yes, stop the optimization and output the phoneme merging rules and new lexicon.

To avoid a computationally heavy exhaustive search through all possible phoneme merges, we limited the search to a constrained space of possible merges. It can be defined according to phonetic knowledge about acoustic similarities between pair of phonemes.

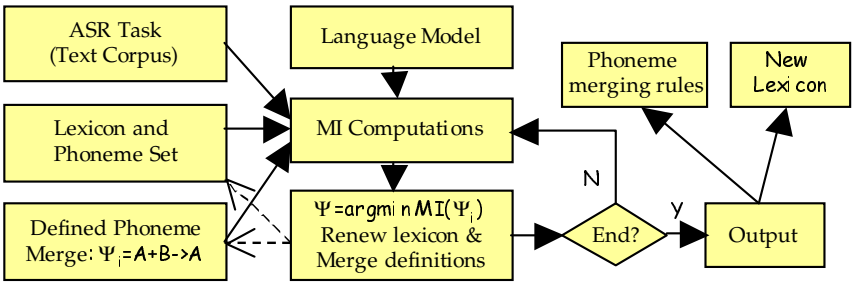

Fig. 2. Illustration of the minimum MI loss based phoneme set reduction.

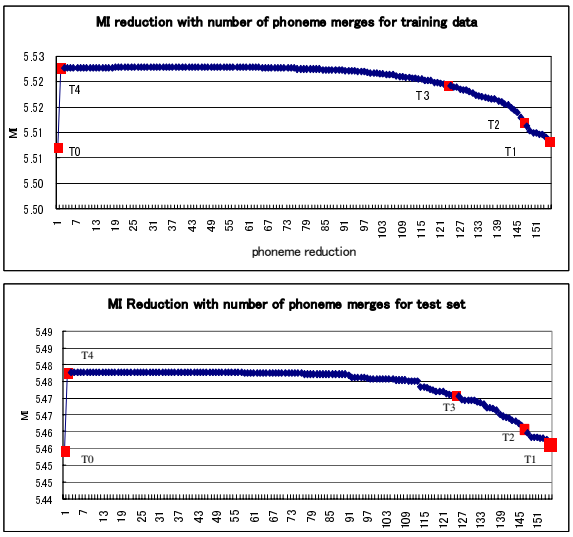

Fig. 3. Illustration of the MI variations with the iterative merges of phonemes. One point move from left-to-right indicates one more merge of two units. The upper panel is for training text corpus, and the lower one for test corpus. The points "T4" stands for the initial 206 phoneme set, whereas those of "T0"s for the non-tonal IFs set.

5. EXPERIMENTS AND RESULTS

5.1. Phoneme Set Design Experimeents

The text corpus (CBTEC) we used is the Chinese version of Basic Travel Conversation Text (BTEC) of ATR. It contains about 200,000 sentences with about one million words. The lexicon size is about 17,000, and the language model is a 2-gram model trained from CBTEC. The size of initial phoneme set is 206 with all tone-dependent units. The initially defined phoneme merges have 433 possibilities, which is designed based on the phonetic similarities of the tonal phonemes. There is one constrain: two different Finals can be merged only when all of them do not have tone-dependencies.

Figure 3 illustrates the MIs of different phoneme sets achieved by merge of units. The figures clearly show that:

- The MI gaps between the points T0 and T4 indicate that some information gets lost when non-tonal IF set is used as the phoneme set for the recognition system.
- There are flat periods of MI variations after T4s in both the training and test data, indicating that a significant number of phoneme merges including tone merges lead to no information loss. Hence, they are the redundancies in the initial full tone-dependent unit set, when given the lexical and language model.
- It offers a flexible way to select a phoneme set to build the speech recognition system according to a balance between the number of units and loss of MI information.

Set	Units	Initials	Finals	Tri-phones
T0	59	21	37	107,441
T1	50	18	31	70,128
T2	59	18	40	114,945
T3	80	20	59	292,651
T4	206	21	184	3,022,775

Table 2. Number of units in the selected sets and the number of their corresponding logical HMMs.

We selected five different unit sets to build our speech recognition systems. T0 is the conventional non-tonal IFs with 59 units; T1 has 50 units and showing a similar MI to that of T0; T2 has the same number of units as T0, but showing a better MI than T0; T3 has 80 units, and showing only slight MI loss from the initial phoneme set T4, which has the full tone-dependent set of 206 units. Table 2 lists the number of Initials, Finals and logical tri-phones in the five ASR systems respectively.

5.2. Speech Recognition Experiments

The training speech data for acoustic models is the Beijing part of ATR Accented Speech (ATRAS). It contains more than 40,000 utterances with a total duration of 43 hours by 96 balanced male and female native Beijing speakers. The test speech data is a subset of CBTEC Putonghua test data. It contains 510 utterances by 5 male and 5 female speakers, each speaker uttering different sentences. The reason to use a slightly accent-mismatched training database is that the ATRAS database has manually annotated phonetic labels, including the tone labels that are different from their lexical forms.

We used the HTK toolkit [11] to build our speech recognition systems to test the five different phoneme sets. The feature vector contains 39 dimensions including standard

MFCC features and log power, together with their first and second ordered derivatives. Cepstral mean subtraction is done at sentence level. We developed phonetic-decision-tree based state-tying tri-phone style HMMs for the different phoneme sets. Each HMM has left-to-right 3 states, the total number of tied states for each model has a similar number of 2,000, and each state has 20 Gaussian mixtures. The speech recognition experiments used the same lexical and language model as those used in phoneme set optimization procedure. The perplexity of the test set is about 40 for the 2-gram language model. The recognition performances are shown in Fig. 4 in Chinese character accuracies.

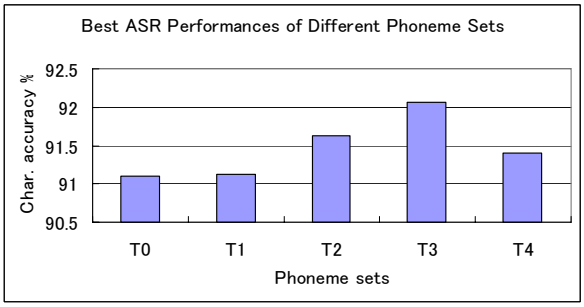

Fig. 4. Character accuracies of speech recognition using different phoneme sets.

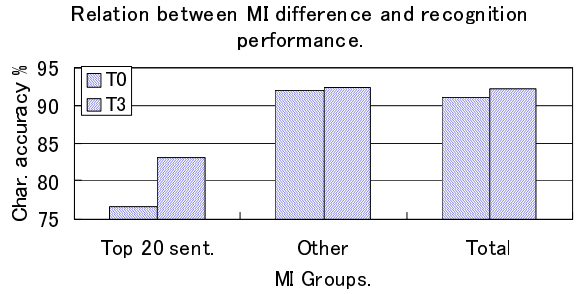

Fig. 5. Illustration of the relationship between MI differences and the recognition performances for T0 and T3. Top 20 represents the 20 sentences with the maximum MI improvements when using T3 instead of T0.

The results showed that

- Almost all the derived unit sets (T1 – T4) showed some better or similar performances compared with the non-tonal set T0, indicating that derived phoneme sets are efficient for the recognition task.
- Although T1 has 9 phonemes less than T0, it still got similar performance to that of T0, indicating the phoneme set more efficient.
- T3 achieved the highest performance, maybe due to its better balance between the number of units and MI information amount than others.
- A close look at the relationship between the MI differences and recognition performances of T0 and T3 separated the 510 test sentences into two groups : one including 20 sentences with the maximum MI improvements, and the one including all left sentences. The first group showed more significant recognition improvements than the other one, as shown in Fig. 5, indicating that positive MI difference is correlated with recognition improvement.

6. CONCLUSION

We presented a novel method of derive compact and efficient tone-dependent phoneme set for building Chinese LVCSR system using MI based criterion. The preliminary experimental results showed the efficiency of the method. The future work will incorporate acoustic confusability measurements into the criterion.

Acknowledgement

This research was supported in part by the National Institute of Information and Communications Technology.

7. REFERENCES

[1] C.-J. Chen and et al., "New methods in continuous Mandarin recognition", Proc. of Eurospeech 1997, Vo. 3, pp.1543-1546.

[2] Ch. Huang and et al., "Segmental Tonal Modeling For Phone Set Design In Mandarin LVCSR", Proc. of ICASSP2004, Vol. 1, pp.901-904.

[3] F. Seide and N. Wang, "Phonetic modeling in the Philips Chinese continuous-speech recognition system", Proc. of Int. Symp. on Chinese Spoken Language Processing, 1998.

[4] C.-J. Chen andet al., "Recognize tone languages using pitch information on the main vowel of each syllable", Proc. of ICASSP, 2001.

[5] Y. Xu, "Production and perception of coarticulated tones", J.A.S.A, **4**, pp. 2240-2253, 1994.

[6] J.-S. Zhang and et al., "Is tone recognition necessary for Chinese speech recognition? ", Proc. of ASJ, Sep. 2002, pp.5-6.

[7] M. Bacchiani and M. Ostendorf, "Using automatically-derived acoustic sub-word units in large vocabulary speech recognition", Proc. of ICSLP, 1998.

[8] D. B. Roe and M. D. Riley, "Prediction of word confusabilities for speech recognition", Proc. of ICSLP, pp.227-230, 1994.

[9] A. Simons, "Predictive assessment for speaker independent isolated word recognisers", Proc. of Eurospeech,pp. 1465-1467, 1995.

[10] D. Torre and et al.,"Automatic alternative transcription generation and vocabulary selection for flexible word recognizers", Proc. of ICASSP, pp.1463-1466, 1997.

[11] S. Young and et al., "HTK Speech Recognition Toolkit ver. 3.2", Cambridge Univ.