TCMSD: A new Chinese Continuous Speech Database [1]

Wang , Dong Zhu, Xiaoyan Wu, Dalei
State Key Lab of Intelligent Tech. & Systems,
Dept. of Computer Science & Technology, Tsinghua University,
Beijing 100084, P.R.China
wd@s1000e.cs.tsinghua.edu.cn

A "good" training and testing database plays a very important role in the establishment of a speech recognition system and in its estimation, where the connotation of "good" includes the coverage of units, their balance, and compatibility with other database. In Chinese speech recognition research, 863CSL(863 continuous speech lib) provides a fairly good standard database for most researchers to build their systems, estimate the validity of new algorithms, and communicate with each other based on general criteria. But some feedback from users shows that 863CSL is not very balance in the sense of syllable, and moreover, since such a database is designed for syllable-based systems, when we wish to extend our system to that of smaller sub-syllable unit based, something it seems inept. For these reasons, we designed a new database, which mainly serves phoneme based recognition systems, and we hope it can give more phoneme coverage rate and balance, especially when cooperates with 863CSL.In this paper, we will discuss some defects we found when we training our gallina system using 863CSL, and give our consideration and strategy to design the new database, TCMSD (Tsinghua Continuous Mandarin Speech Database), especially provide our new method to pick out some "urgent" sentences from huge amount of text materials[1][2], and at last we will report our TCMSD, including it's history, content, and features

Keywords: **speech recognition, database, 863CSL, TCMSD**

1 Introduction

Each researcher in the field of speech recognition will have been impressed by its importance of a "good" training and testing database. A "good" database should provide large unit coverage rate, satisfactory and sufficient utterances for each training unit, a good compatibility with other mature databases, as well as some other aspects, such as noise level (or SNR), broad applicability for more tasks, e.g., gender sensitive systems and noise analysis. In Chinese speech recognition research, 863CSL (863 continuous speech lib) provides a standard to test and estimate the validity of a certain system or a new algorithm, and in fact some systems are also trained using its data [3][4][5], such as our continuous speech recognition system, gallina, but still are there some defects which impel us to design a new database to compensate such inefficiency. In the following section, these defects will be discussed in detail, which is followed by the third section where we will provide our algorithm used to select the "urgent and useful" sentences from a huge text database, and at last we report the process of our TCMSD recording.

2 Some defects of 863CSL applied to continuous speech training

When establish our continuous speech recognition system, gallina, the acoustic model is trained based on 863CSL, and just in this process we noticed that some flaws are there in 863CSL, especially the problem of unbalance of training units, to our system, syllables. In our experiments, we use part of 863CSL with 40 male speakers and 40 female ones, each of which contains one type of sentences, about 520 utterances. Of course, 863CSL works fairly well for our tasks in the whole, and we achieved the

[1] This work is supported by Projects of Development Plan of the State Key Foundation Search (G199803050703)

performance of 61% syllable recognition, but we also find that some syllables can never get enough materials to be trained. After analyzing the database itself, we find that, in our syllable graph of Chinese character, 3.4% syllables have never occurred in the 863CSL database text, and what's more serious is that, we can not get the occurrence of more than 50 times for about 18% syllables, but such a occurrence is necessary for a training process, otherwise, the result models will be digressed to curious status because of data sparseness, even overflow. Table 1 describes the syllable-coverage-rate of 863CSL, from which we can see some syllables too sparse.

TABLE1. Syllable-coverage-rate of 863CSL

N	0	10	30	40	50
Cov	96.6%	90.1%	87.1%	85.9%	82.0%
N	80	150	295	895	6656
Cov	76.7%	67.6%	52.6%	21.9%	0.73%

N: minimum limit of syllable occurrence

Cov: syllable coverage under the limit N

What's more, when we advance our experiments to smaller recognition units, saying, phone and tri-phone, more data sparseness occurs. A statistical experiment shows that, given a hypothesis that each vowel has opportunity to occur before or after anyone of the consonants, the total bi-phone coverage in 863CSL is 58.4%, which includes 63.4% inter bi-phone (bi-phones occur between two Chinese characters) and 44.6% intra bi-phone (bi-phones occur in one Chinese character), and a more sparse tri-phone, which achieve only a coverage of 7.1%. It seems 863CSL is not very apt for sub-syllable based systems, which is one obstacle for us to establish our tri-phone based system- a new version of gallina.

For the above reasons, we designed a new speech database that mainly faces sub-syllable units based systems and, accompanied with 863CSL, gives more bi-phone and tri-phone coverage rate to satisfy the training tasks.

3 Sentence selection strategy for TCMSD

Indeed, although some insufficiency there when applied sub-syllable training, 863CSL is a fairly good standard data resource even for training, and so, without any reason to establish the new database without any consideration of the current valuable corpus, we design TCMSD, from the very beginning, with a clear purpose to compensate instead of to re-build. As the same time, we hope to select a set of sentences that may cover as many syllable, bi-phone and tri-phone as possible but still restrain it in a small set size. These two objects direct us to assume a rational criterion to estimate our future database, and then based on that criterion, we can decide which sentence should be included while others should be rejected.

Our strategy is choosing serials of sentences from a large txt database to maximize an evaluation formulation. In fact, we use a text database as the "data-pool" which even contains up to 3 million sentences.

Initially, we set a set size to limit the total sentences that will be selected, for instance, 1000, and then the first 1000 sentences in 863CSL are included in the selection set which forms an initial database, with other sentences in 863CSL and other 3 million ones used as candidate sentences.

At first, we set a score for a new coming sentence, which includes all contributions of each phoneme in that sentence. For each phoneme, we hypothesize three types of contribution scores, reflecting the uni-phone, bi-phone and tri-phone respectively and for their different emphasis, we can give each type different weight in the whole contribution. In our design, we incline more weight for tri-phone, and so the weights are set as 0.2, 0.3,0.5.

To design a proper formulation to estimate its contribution when a uni-phone or tri-phone is added to the original database, we should consider an obvious fact that sparser the phone is, higher score the phone should offer. Based on such consideration, we use the following form to estimate the contribution score.

$$s = \frac{1}{((aI)^2 + \beta)^r} \qquad (1)$$

Where s is the score contributed by a uni-phone, bi-phone or tri-phone, a, a scale factor to lengthen x-axle, β, a tail value to prevent a zero-value denominator, r, a power to control the gradient of the estimating function, and I is the number of such a phone contained in the database by now. Formulation (1) is flexible by adapting its parameters to discriminate uni-phones, bi-phones and tri-phones, and if a new sentence comes, each phone in it may contribute to the database a uni-phone, a bi-phone and a tri-phone, which leads to the estimating formulation for each phone as following,

$$S = \sum_N w_n s_n \qquad (2)$$

Where S refer to the whole score a new phone offers to the by-now database, and w_n the weight of sub-phone s_n whose footnote n is an order indication from 1 (referring to uni-phone) to the top-most order N, which is 3 in our case. As mentioned above,

$w_1 = 0.2$, $w_2 = 0.3$, $w_3 = 0.5$, and for a more sensitive bi-phone score function, we set $a_2 > a_3$, which, 10 and 30 respectively in (1). For a new sentence, only we need to do is to accumulate each S for every phone in it and then we can get a whole estimation score for that sentence.

Our sentence selection strategy is based on the contribution scores discussed above, and in the database a sentence with larger score will replace that with a smaller one. As is mentioned, we first put the front 1000 sentence in 863CSL to the database, and the remained sentences will be looked up one by one to decide whether or not the sentence should be selected in under our rule. By comparing the total score increase if we insert a certain sentence A and the total score decrease if we delete a sentence B in the current database, we get a score increase if replace B with A, and such process will continue until all the sentences in the database have been compared with A, which will show a most "profit" replacement and that is just which we want. After all the candidate sentences being looked up, a new database is achieved which maximize our estimation formulation.

Using our selection algorithm, we get our TCMSD database of 1000 sentences whose tri-phone cover rate is 14.3% and bi-phone cover rate 71.5% by itself, while the corresponding values are only 58.4% and 7.1% in 863CSL with its 1500 sentences. This result also demonstrates the effectiveness of our selecting method. We also tried a larger size, but the result shows that 1000 is a good tradeoff between sentence amount and coverage achievement. Table 2 describes more details of our new database-TCMSD.

TABLE 2. Comparison of Unit coverage

DATABASE	863CSL	TCMSD	863CSL +TCMSD
Sentence	1500	1000	2500
Bi-phone	58.4%	71.5%	73.4%
Tri-phone	7.1%	14.3%	16.8%

4 Main feature of TCMSD

To be compatible with 863CSL, we also select a mono-channel PCM sample under the sample rate 16000Hz to record TCMSD, and also we apply a group-division strategy to organize the new database. There are total four groups A, B, C and D, with 250 sentences in each group, and each speaker is required to record at least one group. The difference from 863CSL is that we don't deliberately seek very silent recording condition and very fluent speaker, because we found in our experiments that too pure training material doesn't benefit the practical application, and so that we hope by collecting more sentences under natural speaking condition and by natural speaker, we can get our more natural models at least in the aspect of database.
By now, totally 30 sets (120 groups) have been recorded and collated, and we will finish another 70 sets by the end of this year. We hope the new database will benefit the Chinese speech research, especially for such systems based on sub-syllable units but being baffled by data sparseness.

5 Conclusions

In this paper, we describe our design of a new continuous Chinese speech database, TCMSD, presenting its main feature and our process for its recording. We hope and believe this new database will give much help to the researchers who is troubled by data sparseness, looking for more natural training materials, and seeking more robust acoustic models.

References

[1] Price, Patti; Fisher, William M., .etc, "DARPA 1000-WORD resource management database for continuous speech recognition ", ICASSP88, pp. 651-654.

[2] Zue, Victor,; Seneff, Seneff, .etc, " Speech database development at MIT. TIMIT and beyond", Speech Communication ,Vol. 9, No.4, pp. 351-356.

[3] Huang, Hank Chang-Han; Seide, Frank , "Pitch tracking and tone features for Mandarin speech recognition", ICASSP 2000, pp. 1523-1526(2000).

[4] Liu, MingKuan; Xu,Bo, etc., " Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling", ICASSP 2000, pp. 1025-1028(2000).

[5] Song,Zhanjiang; Zheng,fang, .etc, "Statistical knowledge based frame synchronous search strategies in continuous speech recognition", ICASSP2000, pp. 1583-1586(2000).