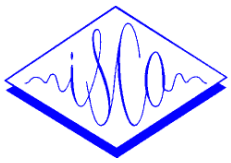

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

The HKU-USTC Speech Corpus

Chorkin Chan

Department of Computer Science,
University of Hong Kong

Ren-hua Wang

Department of Radio Electronics,
University of Science and Technology of China

ABSTRACT

A design of spoken Chinese corpus is proposed which consists of five sub-corpora C1 to C5. The design principles are (1) Mono-syllables are important not only for the recognition of isolated syllables but also for the recognition of connected spoken Chinese because they simplify the isolation of phonetic information in the training phase, (2) A corpus including all inter-syllable triphones captures all the immediate left- and right- context of phones at syllabic boundaries. This is a logical and practical compromise between exhausting all possible syllabic transitions and keeping the corpus building effort at a manageable level. C1 consists of 433 mono-syllables and 4 consonant clusters in each of its four versions. Each toned syllable exists in at least one of the four versions and the 433 mono-syllables in each version include all the syllables in one of the four regular tones (yin, yang, shang and qu) plus all the neutral-toned ones. C2 is a collection of 16 digit strings each ranging from 4 to 7 digits in length. These strings exhaust all the inter-digits triphones. C3 has 30 geographic names each of two to five syllables. C4 consists of 859 short phrases each of six to nine syllables long forming the bulk of the inter-syllable triphone collection. C5 is a catch-all sub-corpus composed of all the inter-syllable triphones which do not appear in C2 to C4. These triphones are very seldomly used in the Chinese language today and can be ignored in recognizer training for practical purposes.

Keywords:-Speech database, isolated speech, connected speech, Putonghua, inter-syllable triphones, coarticulation between syllables

1. Introduction

Despite recent interest in analyzing and recognizing spoken Chinese commonly known as Putonghua or Mandarin, no systematic effort in establishing a comprehensive speech corpus involving a large number of speakers with various accents, age groups and sexes has been reported. Such a corpus has great importance for the following reasons:

1. A pure speaker dependent speech recognizer requires too much training by each speaker. It is much more efficient to have the recognizer trained with some universal speech data and then fine tuned with speaker dependent data to adapt to the characteristics of the individual speaker. A lot of research had been conducted on adapting speaker dependent recognizers to suit new speakers but there are good reasons to believe that such systems will perform less satisfactorily than a speaker independent recognizer with adaptation.
2. Constructing a large speech corpus is expensive and it requires proper expertise and equipment and that explains the absence of a Chinese speech corpus similar in nature and size to TIMIT. As a consequence, speech researchers of the Chinese language must live with small scaled, home-made and often unprofessionally designed and built corpora. The drawback is extremely significant. Without a common test-bed, it is difficult to compare one's research results against that of others. Secondly, without a database of international status, researchers and developers of spoken Chinese recognizers are handicapped severely to compete with their colleagues of other countries. After all, despite the fact that the recognition techniques used should be language independent, a recognizer cannot avoid, to some degree, having the recognition strategy specially designed to suit the target language.

Chen et al [1] had proposed a Chinese language speech database. Unfortunately, they did not justify the design with any quantified objectives nor statistics making the suitability of the database for training a recognizer of connectedly spoken Chinese questionable. In fact, certain syllabic transitions appear redundantly while some transitions simply do not exist in the database at all without any explanation. According to these authors, the database was designed for analysis, data compression, synthesis and recognition of the spoken Chinese language. However, the requirements for these purposes can be vastly different from each other and there is really no need to have one design for all objectives.

The designers of TIMIT [2] [3] stressed the importance of capturing coarticulation in connected speech. This is certainly an important aspect of speech research in any language. However, for the Chinese language with mono-syllabic characters as building units, including all the isolated syllables in the corpus has a special significance. First, it eases the problem of phonetic labeling which is important in the training phase of a recognizer. Secondly, there is still a large number of researchers interested in recognizing isolated Chinese syllables because it's relative simple to construct such recognizers. Corpora like TIMIT were designed to cover the left- and right- context of each phone as much as

possible because any attempt for an exhaustive coverage is highly impractical considering the amount of data to be collected that demands.

Kurematsu et al [4] and Shirai et al [5] reported the Japanese efforts in building spoken Japanese corpora. They did include a certain amount of isolated words besides connected speech. Again, no systematic design principle to capture the coarticulation effects was mentioned.

/a*/	/a+/	/b/	/c/	/ch/	/d/	/e/	/ei/
/en/	/eng/	/er/	/f/	/g/	/h/	/j/	/k/
/l/	/m/	/n/	/o/	/ou/	/p/	/q/	/r/
/s/	/sh/	/t/	/w/	/x/	/y/	/z/	/zh/

Table 1 - 32 first-phones

2. The Design Principles

The purpose of building a speaker indepenedent speech recognizer is not only to be able to cope with the average speakers but also to be more adaptable to an arbitray speaker. This is because the training speech corpus is so designed that hopefully, someone else's voice with similar characteristics had already been captured so that if a suitable adaptation strategy is employed, good performance can still be expected. To this end, the corpus must consists of all possible allophones of the language spoken by a large number of speakers. A second consideration is to capture all the coarticulation effects one can expect in the language. Furthermore, the corpus must be designed to suit many research objectives in order to make an impact on the research community. On the one hand, one wants the corpus to encompass all speech information from isolated syllables as well as between syllable coarticulation effects [6]. On the other hand, one wants as little redundant information as possible in the corpus to make it concise and manageable and in particular, one wants it flexible enough so that one can have it tailored for one's purposes. For these reasons, the corpus is divided into five components C1 to C5 so that researchers have the freedom of selecting a combination of these components to suit their specific needs. C1 consists of essentially all the isolated syllables while C2 to C4 consist of phrases which are not necessarily syntactically and semantically perfect. C5 is a collection of syllabic pairs the inter-syllable triphones of which are not included in C2 to C4. These phrases and syllabic pairs will exhaust all left- and right- context of all phones at syllabic boundaries. On the one hand, it is important that these phrases are syntactically and semantically sound so that speakers can read them naturally. On the other hand, since one wants to minimize the size of the corpus by avoiding redundant syllabic combinations as much as possible, some phrases will necessarily be nonsensical and grammatically less than perfect. Besides, the Chinese language is such that mono-syllabic characters are building units with which names and new terms can be constructed rather freely and no one can say what syllabic combinations will never be acceptable as a new term or jargon. There is a total of 404 syllables according to [7]. Each syllable can be read in one to five different tones which are variations of the time pattern of pitch. In order to capture all possible syllabic transitions, one is talking about 404^2=160K syllabic pairs which are simply impractical to handle. However, a syllable may start with any one of 32 phonemes and may end with any one of 12 phonemes. If one can be satisfied with only diphones (two neighbouring phones) over syllabic transitions, there are only 384 syllabic transitions to be included into the corpus besides the mono-syllables. Unfortunately, a corpus of such syllabic pairs will provide incomplete left- and right- context of phones at syllable boundaries. To meet the stated design principles, one must consider in terms of inter-syllable triphones (three phones in succession). In order to understand what must be done, it is necessary to first enumerate all the first-phones, first-diphones, last-diphones and last-phones of Chinese syllables. Employing the Pin-yin symbols with some minor modifications to allow entering phonetic information using a keyboard, the 32 first-phones of Chinese syllables are listed in Table 1.

/a/	/ai/	/an/	/ang/	/ao/	/ba*/	/ba+/	/bei/
/ben/	/beng/	/bi*/	/bo/	/bu/	/ca*/	/ca+/	/ce/
/cen/	/ceng/	/ci/	/cou/	/cu*/	/cha*/	/cha+/	/che/
/chen/	/cheng/	/chi/	/chou/	/chu*/	/da*/	/da+/	/de/
/dei/	/deng/	/di*/	/dou/	/du*/	/e/	/ei/	/en/
/eng/	/er/	/fa*/	/fan/	/fei/	/fen/	/feng/	/fi/
/fo/	/fou/	/fu/	/ga*/	/ga+/	/ge/	/gei/	/gen/
/geng/	/gou/	/gu*/	/ha*/	/ha+/	/he/	/hei/	/hen/
/heng/	/hou/	/hu*/	/ji*/	/ju*/	/ka*/	/ka+/	/ke/
/kei/	/ken/	/keng/	/kou/	/ku*/	/la*/	/la+/	/le/
/lei/	/leng/	/li*/	/lo/	/lou/	/lu*/	/lv*/	/ma*/
/ma+/	/me/	/mei/	/men/	/meng/	/mi*/	/mo/	/mou/
/mu/	/na*/	/na+/	/ne/	/nei/	/nen/	/neng/	/ni*/
/nou/	/nu*/	/nv*/	/o/	/ou/	/pa*/	/pa+/	/pei/
/pen/	/peng/	/pi*/	/po/	/pou/	/pu/	/qi*/	/qu*/
/ra*/	/ran/	/re/	/ren/	/reng/	/ri/	/rou/	/ru*/
/sa*/	/sa+/	/se/	/sen/	/seng/	/si/	/sou/	/su*/
/sha*/	/sha+/	/she/	/shei/	/shen/	/sheng/	/shi/	/shou/
/shu*/	/ta*/	/ta+/	/te/	/tei/	/teng/	/ti/	/tou/
/tu*/	/wa*/	/wa+/	/wei/	/wen/	/weng/	/wo/	/wu/
/xi*/	/xu*/	/ya*/	/yan/	/ye/	/yi*/	/yo/	/you/
/yong/	/yu*/	/za*/	/za+/	/ze/	/zei/	/zen/	/zeng/
/zi/	/zou/	/zu*/	/zha*/	/zha+/	/zhe/	/zhei/	/zhen/
/zheng/	/zhi/	/zhou/	/zhu/				

Table 2 - 188 first-diphones

/a/	/ba/	/ca/	/cha/	/da/	/fa/	/ga/	/ha/
/*ia/	/ka/	/la/	/ma/	/na/	/pa/	/sa/	/sha/
/ta/	/*ua/	/za/	/zha/	/e/	/ce/	/che/	/de/
/ge/	/he/	/*ie/	/ke/	/le/	/me/	/ne/	/re/
/se/	/she/	/te/	/*ue/	/ze/	/zhe/	/*ang/	/*eng/
/*ing/	/*ong/	/ai/	/bi/	/ci/	/chi/	/di/	/*ei/
/ji/	/li/	/mi/	/ni/	/pi/	/qi/	/ri/	/si/
/shi/	/ti/	/xi/	/yi/	/zi/	/zhi/	/*an/	/*ian/
/*en/	/*in/	/*vn/	/o/	/*ao/	/bo/	/fo/	/lo/
/mo/	/po/	/*uo/	/yo/	/er/	/bu/	/cu/	/chu/
/du/	/fu/	/gu/	/hu/	/*iu/	/ku/	/lu/	/mu/
/nu/	/*ou/	/pu/	/ru/	/su/	/shu/	/tu/	/wu/
/zu/	/zhu/	/ju/	/lv/	/nv/	/qu/	/xu/	/yu/

Table 3 - 104 last-diphones

Here the notation /a*/ stands for the first-phone in syllables "a","ang" and "ao" while /a+/ stands for that in syllables "ai" and "an". /en/, /eng/ and /ou/ stand for the first-phones in syllables "en", "eng" and "ou" respectively.

The first-diphone of a syllable is the concatenation of the first two phones of the syllable. For example, /y//a/ is the first-diphone of the syllable "ya". For single phoneme syllables like "a", "e", "er" and "o", their diphones are defined as concatenations of their respective single phonemes. There are altogether 188 syllabic first-diphones. In Table 2, we use the letter v to stand for the umlau u and /αβγ / to stand for the first-diphone of syllable "αβγ" where α, β and γ are generic letters meaning that they can stand for any letter for any length. The following exceptions are to be observed though:

/αβγ+/	standing for the first-diphone of syllables "αβγn" and "αβγi", e.g., /cha+/ for the first-diphone of "chai" and "chan".
/αβγ*/	standing for the first-diphone of syllables starting with letters αβγ except " αβγi" and " αβγn", e.g. /zha*/ for the first-diphone of "zha", "zhang" and "zhao".
/ αβu*/	meaning / αβγ*/ as above except that γ must be the letter u and it also includes the diphone of " αβong", e.g., /cu*/ for the first-diphone of "cong", "cu", "cui", "cun" and "cuo"

Similarly, Tables 3 and 4 list respectively 104 last-diphones and 12 last-phones that can be found in Chinese syllables. Here, we assume again that syllables "a", "e", "er" and "o" have last-diphones just the same as their corresponding first-diphones. The notations used in these tables are such that:

/αβγ/	is the last-diphone of syllable "αβγ" where α, β and γ are generic letters.
/*αβγ/	is the last-diphone of any syllable ending in αβγ, e.g. /*in/ for that of "bin", "chin", "lin",.... There are some exceptions though. The last-diphone of "yan" is denoted as /*ian/, that of "ya" as /*ia/ and that of "ye" as /*ie/. /*ei/ stands for the last-diphone of syllables like "bei", "dei",... and "cui", "chui", "dui", ... also. Likewise, /*en/ stands for the last-diphone of syllables like "ben", "den",... and "cun", "chun", "dun", ... also.
/ie/	is the last-phone of "ye"
/i1/	is the last-phone of "bi"
/i2/	is the last-phone of "ci"
/i3/	is the last-phone of "chi"

Each inter-syllable triphone is a combination of (last-phone, first-diphone) or (last-diphone, first-phone) occurred in a syllable-to-syllable transition with the last-phone or last-diphone derived from the syllable on the left and the first-diphone or first-phone derived from the syllable on the right. There is a total of (12 x 188 + 32 x 104 =) 5584 possible inter-syllable triphones with some overlaps. In general, one cannot avoid having inter-syllable triphones appearing in the phrases more than once at times. For example, let α, β, γ, δ, ε and η be phones. Assume in the language that there are only three syllables S_1, S_2 and S_3 which have their last phones equal to δ while the last-diphones of them are αδ, βδ and γδ respectively. Assume also that there is only one syllable S_4 the first-phone of which is ε. The first-diphone of this syllable is εη. In order to exhaust all (last-diphone, first-phone) combinations, S_1, S_2 and S_3 must pair up with S_4 thus producing two redundant inter-syllable triphones of (δ, εη). Intra-syllable triphones are already included in the corpus because of the complete collection of isolated syllables in C1.

3. Sub-Corpus C1

Basically, Chinese syllables can be articulated in four different tones (yin, yang, shang and qu respectively). Some

/a/	/e/	/ie/	/er/	/i1/	/i2/	/i3/	/n/
/ng/	/o/	/u/	/v/				

Table 4 - 12 last-phones

syllable-tone combinations may not be actually used in speech and correspond to no Chinese characters. The other combinations that actually correspond to currently used Chinese characters are hereafter referred to as valid toned syllables. In addition, 29 syllables can appear in the neutral-tone (the light or fifth tone) also. Furthermore, according to [7], there are 4 consonant clusters, viz., "hm", "hng", "m" and "ng" which can be articulated and they actually have Chinese characters associated with them. C1 appears in four versions. Each version contains each syllable at least once, together with all the 29 neutral-toned syllables and the 4 consonant clusters. Each valid toned syllable appears in at least one of the four versions. As a result, each version has a total of 433 syllables plus 4 consonant clusters with roughly equal number of syllables in each of the four regular tones. These four versions of C1 are tabulated in Appendix I in which each toned syllable is listed with an associated Chinese character to help the speakers reading them out.

The idea of structuring C1 in this manner is that it may be too much to ask a speaker to read all the syllables in all tones. However, one doesn't want to leave out any valid toned syllable. As a compromise, each speaker will read one version of C1 only and will cover every syllable in at least one tone. If each version is read by roughly the same number of speakers, each toned syllable has approximately the same chance of being read. The number of neutral-toned syllables and consonant clusters are small hence it is wise to have them read by every speaker. Asking a reader to read in the same tone for all syllables will be too monotonous, hence the syllable-tone combinations are somewhat randomized in each version. Since some syllables do not appear in all tones, some syllable-tone combinations appear more than once to make up the equal sizes of the four versions.

4. Sub-Corpus C2

A recognizer of connected digit strings will find many applications. The design of C2 is such that it includes all possible inter-digit triphones in as few strings as possible. Each string cannot be too long otherwise the speakers will find them difficult to read naturally. There is a total of 16 such strings of six digits each except three of them which have four digits, five digits and seven digits respectively. In Appendix II, they are listed together with their numeral forms. Notice that "1" is read as "yao" and not as "yi" for ease of recognition in a noisy environment.

5. Sub-Corpus C3

C3 is composed of 30 geographical names with 2 to 5 syllables each. The purpose of setting up C3 as a separate subcorpus is to provide a small database of connected speech in terms of isolated words (names) which will be very useful for testing speech models of multi-syllable isolated words. These names can be found in Appendix III.

6. Sub-Corpus C4

C4 consists of 859 short speech phrases. It would be very

	neutral	yin	yang	shang	qu
neutral	4	42	50	42	62
yin	72	241	315	174	324
yang	76	293	356	371	409
shang	46	201	248	10	316
qu	97	307	459	261	527

Table 5 - Tone transition statistics

tiring for any speaker speaking all (859 + 16 + 30 =) 905 phrases from the three sub-corpora C2, C3 and C4. It is therefore suggested that each speaker speaks only the phrases in C2 and C3 plus a few tens of phrases selected randomly from C4. Care must be taken that each phrase in C4 is covered by approximately the same number of speakers with a uniform distribution over different age groups, sexes and accents. In Appendix IV, each phrase is listed together with the associated string of Chinese characters to help the speakers read them naturally.

Table 5 provides statistics of tone transitions which are the number of times syllables in one tone transiting to the next syllable of another tone within the phrases in C2, C3 and C4. These are the actual tones spoken in natural speech rather than the tones the characters are associated with in isolation, i.e., tone rules have been applied already.

7. Sub-Corpus C5

Some inter-syllable triphones involving the syllables "a", "ai", "an", "ang", "ao", "den", "ei", "en", "eng", "fiao", "lo", "nou", "o", "ou" and "yo" are missing in sub-corpora C2 to C4. C5 is a collection of 578 inter-syllable triphones needed to complement the triphone collection in C2 to C4. These triphones do not commonly appear in the language hence for practical purposes, one can ignore C5 and live with a sub-optimal but practical speech recognizer training corpus. Appendix V gives an account of these missing triphones.

8. Conclusion

A design of a spoken Chinese corpus has been proposed. The design is certainly not optimal but in the opinion of the author, this is a practical and comprehensive design encompassing the most important speech information for recognizer construction. After all, given the ambiguity of syntactic and semantic soundness, one cannot help wondering whether there is such a thing as an optimal design for corpus of this nature.

In view of the fact that this is a conference paper subject to strict limitation on its length, all appendices will be skipped but they can be made available from the first author of this paper upon request.

9. Acknowledgement

The author is grateful to Mr. P.K. Wong for his help in typing the Chinese text.

10. References

[1] X.Q. Chen, C.L. Li, F.Y. Mo and S.N. Lu; "A Chinese Language Speech Database", Proceedings of Conference on Speech, Communication and Image Processing, pp. 127-129, 1987.

[2] L.F. Lamel, R.H. Kassel and S. Seneff; "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus", Proc. Speech Recognition Workshop, pp. 100-110, 1986 (DARPA).

[3] J.S. Garofolo, "The Structure and Format of the DARPA TIMIT CD-ROM Prototype", Documentation of DARPA TIMIT.

[4] A. Kurematsu, K. Takeda, H. Kuwabara and K. Shikano, "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis", Proc. of ESCA Workshop, pp. 2.3.1- 2.3.4, 1989.

[5] K. Shirai, H. Fujisaki and S. Itahashi, "Speech Database Projects in Japan --- Present and Future ---", Proc. of ECSA Workshop, pp. 2.4.1 - 2.4.4, 1989.

[6] Mei-yuh Hwang, Hsiao-wuen Hon and K.F. Lee, "Modeling Between-Word Coarticulation in Continuous Speech Recognition", Proc. Eurospeech 89, pp.5-8, Paris, 1989.

[7] Xiandai Hanyu Cidian, Commercial Press.