# Sinica Continuous Speech Prosody Corpora (COSPRO) and Toolkit

## ● Overview: Research orientation and theoretical background

The Sinica COSPRO (Mandarin Continuous Speech Prosody Corpora) and Toolkit is designed, collected and annotated by Dr. Chiu-yu Tseng and her research group at the Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei, Taiwan. The package of 4 DVD's contains 10.5 GB (7.7 GB annotated) of speech corpora and the Toolkit. Funding resources for corpus collection and toolkit development came exclusively from Academia Sinica, mainly under the support of three Academia Sinica interdisciplinary Theme Projects, namely, "Collaborating Researches on Chinese Information Processing-Subproject on Mandarin Chinese Speech Database (1994.7-1999.7)", "Knowledge Representation and Language Engineering for Mandarin Chinese --- Man-machine Voice Interface Environment and Its Tools (1997.7—2002.6)" and "New Directions for Mandarin Speech Synthesis : From Prosodic Organization to More Natural Output (January 2003—December 2005).

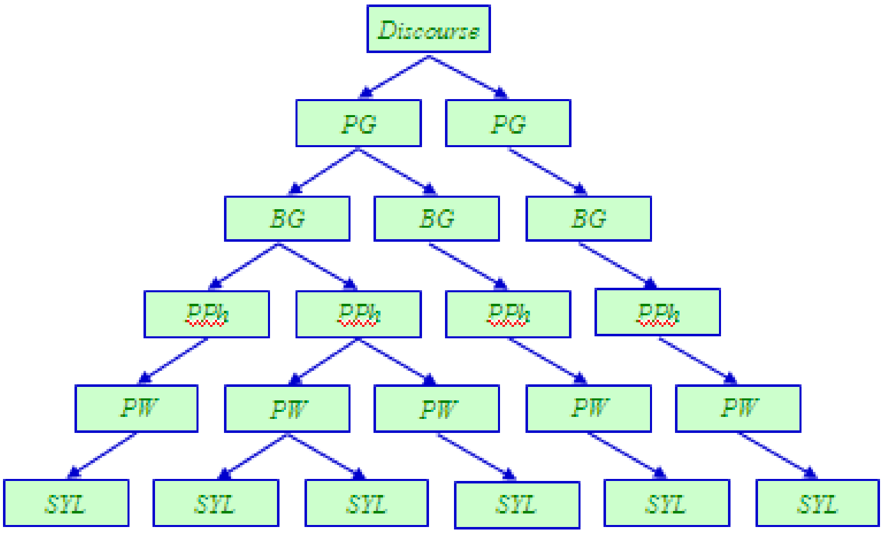
The major feature of the corpora and toolkit is its linguistic orientation and theoretical implications, namely, discourse or narrative effects exhibited in fluent speech prosody. We believe that any attempt to derive or simulate prosody of continuous running speech, especially when adopting a corpus approach, must account for how speech flows continuously. In particular, the speech units on which fluent speech prosody are planned during on-line speech production, and possible prosody effects in speech perception. In other words, factors such as planning threshold, planning strategies, prosodic units, boundary information and boundary breaks that collectively make up the melodic and rhythmic structures and patterns and cause speech flow sound continuous. The question is: how these above-mentioned factors, together with syntactic structures, semantic interpretations and speakers' intensions, collectively contribute to the communication infrastructure of fluent speech? What are the necessary expressions in speech delivered via prosody? Due to interactions among these factors, we view fluent speech as a mixture of both quantal and slurred acoustic signals that can not and should not be viewed as concatenation of unrelated prosodic units, be they large or small. Hence the focal points of our research orientation are how this mixture should be studied through speech corpora and how prosody should be analyzed. In addition, we also believe the collective and integrated effects from all of these factors also form the basis for on-line speech processing. Therefore, how and in what kind of units fluent speech is perceived by listeners is just as important as analyzing the physical speech signals in relatively large and/or small units. We need to keep in mind that the speech chain can only be formed by linking production to perception.

That is why we chose to collect narrative speech, analyze the corpora from a top-down perspective on perceived units and boundaries across running speech and came up with a speech unit of multiple-phrase complex sentences or speech paragraphs instead of individual phrases. We thus postulated a speech unit called Prosodic Phrase Group (PG) not referred to in traditional phonetic research as well as a corresponding hierarchical multiple-phrase prosody framework. Through quantitative analyses of the corpora in COSPRO, we were able to obtain systematic cross-phrase evidences in every acoustic domain from each prosodic level, namely, global F0 contour patterns, syllable duration adjustment, intensity distribution patterns and boundary break patterns. Furthermore, we were also able to account for how these layered contributions cumulatively formed the overall prosody output of multiple-phrase speech paragraphs in narratives.

Two major characteristics distinguish our top-down hierarchical PG framework from orientations of other prosody analyses. One is the units and boundaries in fluent speech under investigation; the other is how to implement systematic scientific findings to technological applications. On the units and boundaries in fluent speech, two acoustic domains received special attention: one is the role of phrasal intonation that constitutes the overall fluency melody and another temporal allocation across phrases that constitutes the overall speech rhythm. We found that individual phrasal intonations are no longer unrelated independent prosody units in fluent speech, but rather, sister constituents subject to higher commands from PG. As a result, PG specified modifications are necessary. That is, PG-specified positions require the phrases under grouping to adjust their respective contour patterns in order to signal the beginning, continuation and termination of a speech paragraphs. On fluent speech melody, we found that cross-phrase cadence templates exists, thus explaining why simple concatenation of unrelated individual intonations strings will not yield fluent speech melody. These PG-specifications are not always syntactic constraints only; they could also be semantic or speaker-intended, thus making investigation of fluent speech prosody more complicated. On fluent speech rhythm, we found consistent cross-speaker and cross-phrase syllable duration templates and temporal allocation patterns at each and every prosodic level, thus explaining how fluent speech rhythm is formed. Cross-phrase rhythm cadence templates also exist, requiring individual syllable durations to modify in accordance with each prosodic level, and cumulatively contributes to the final and overall rhythm and beat of fluent speech as did the contour patterns. In other words, higher-up commands from speech paragraph cause cross-phrase F0 and duration patterns to adjust and modify systematically. To yield fluent speech prosody, intonation modifications alone are insufficient unless syllable durations are also modified; both modifications are in accordance with respective cross-phrase templates in each and every prosodic layer. Note that our findings in fluent speech temporal allocations and rhythmic outputs are significant in at least the following three aspects: one is how Mandarin Chinese syllables should be studied in fluent speech; two is how speech rhythm could be investigated, and finally how we can no longer focus on F0 contours only and reason we have pretty much done most of the work for prosody.

To implement our findings to speech technology development, a correlative modular acoustic model was also constructed. The model can be used to manipulate F0 contours, syllable durations, intensity distribution and boundary breaks independently or collectively, and is ready to be used with any synthesis program for prosody adjustments. However, we will continue to work on boundary information as well as other expressions in speech in the future.

The prosody framework stresses a hierarchical governing effect that groups phrases into speech paragraphs most notably exemplified in narratives, and specifies cross-phrase relationship in each and every of the acoustic domains involved. Layered prosodic contributions from different levels of the hierarchy that cumulatively constitute overall fluent speech prosody. In Tseng's (2005) framework, prosodic layers are specified and phrase intonations are required to adjust by their respective positions within a PG; cross-phrase F0 contour cadence templates, syllable duration cadence templates, intensity distribution patterns and corresponding boundary patterns are derived. The following figure is a schematic representation of the framework.

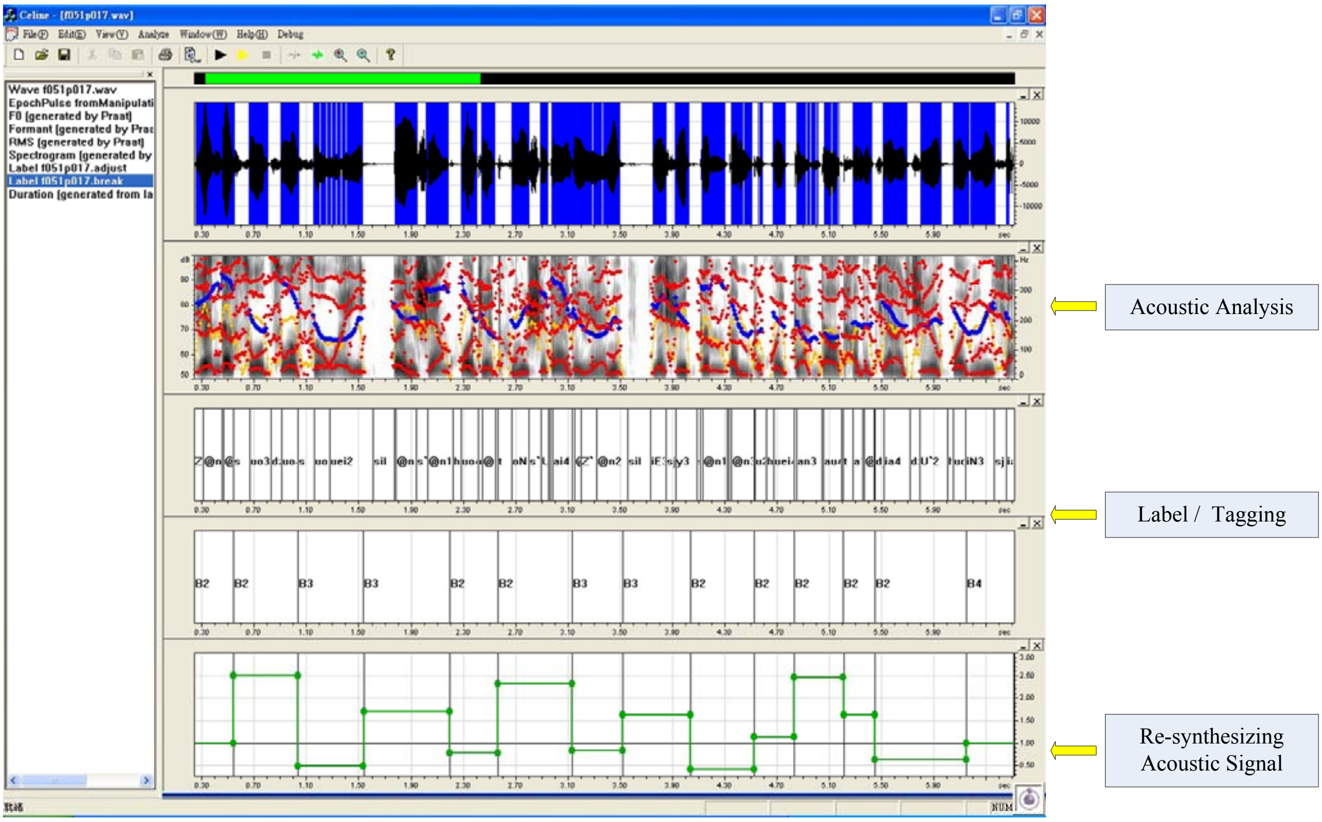


Various kinds of fluent speech data were designed over time to bring out different properties of fluent speech prosody that no short and isolated sentence would yield. The corpora included a total of 10.5GB of recorded speech. Almost all of the corpora are read discourses, with only 76MB in spontaneous narratives. The Toolkit is a perceptually based annotation platform catered to label perceived boundary breaks and to derive various levels of prosodic units in fluent speech.

There are 9 sets of speech corpora, namely, (1.) Phonetically Balanced Speech Database (COSPRO 01, 1.99GB, 18:38 recording time), (2.) Multiple Speaker Speech Corpus (COSPRO 02, 2.08GB, 19:29 recording time), (3.) Intonation Balanced Speech Corpus (COSPRO 03, 2.38GB, 31:10 recording time), (4.) Stress-pattern Balanced Speech Corpus (COSPRO 04, 243MB, 48m recording time), (5.) Lexically-balanced Speech Corpus (COSPRO 05, 575MB, 35:50 recording time), (6.) Focus-balanced Prosody Group Speech Corpus (COSPRO 06, 759MB, 7:30 recording time), (7.) Text-type/Speaking-style Varied Speech Corpus (COSPRO 07, 577MB, 1:32 recording time), (8.) Prosody Balanced Monosyllable Corpus (COSPRO 08, 1.9GB, 15:12 recording time), and (9.) Comparable Spontaneous/Read Speech Corpus (COSPRO 09, 76MB, 42m recording time). Each corpus was designed to bring out different prosody features involved in fluent speech.

Each set of speech database consists of processed and unprocessed speech data. The speech data were collected according to the following procedures: (1.) designing text pieces with specific prosody features, (2.) recruiting appropriate speakers and (3.) recording speech data in sound-proof chambers at sampling rate of 16000Hz and in 1-channel 16-bit linear format in sound-proof chambers into waveform files (*.wav). Processed speech data involved the following procedures: (1.) checking recorded speech with corresponding text, editing sound files that are too long into shorter pieces and subsequently editing text to match file size, (2.) converting text into SAMPA files (*SAMPA), (3.) performing broad transcription using the HTK toolkit (*phn), (4.) performing human spot checking for correct segments and hand-adjust segmental boundaries (*adjusted), (5.) labeling prosodic boundaries (*break) by human transcribers and checked for intra- and inter-transcriber consistencies, and (6.) analyzing prosodic units and features to test Tseng's prosodic hypothesis and framework. More comprehensive information of the fluent prosody framework and modeling is available in a paper entitled "Fluent Speech Prosody: Framework and Modeling" by Tseng et al (2005).

The Sinica COSPRO Toolkit is a platform of analysis and tagging tool that process data into prosodic units. It can be used with the modular mathematical model that makes use of PRATT synthesis to manipulate each of the above 4 acoustic domain either separately or collectively to synthesize overall prosody. The Toolkit is a window-based platform that consists of three basic functions: (1.) speech analysis functions that include basic acoustic analysis such as sound wave, point process, F0 tracking, intensity, formants and spectrographs; (2.) labeling functions that label speech files into phonemes using SAMPA notations and tag perceived boundary breaks using break indices from 1 to 5; (3.) Re-synthesis functions that include two blocks at the current stage, the F0 block and the syllable duration block that allow manipulations of the two acoustic features specified independently or collectively. Multiple windows can be simultaneously opened at each function for display and analysis. The Toolkit also features analysis functions with a simpler interface so that visual displays are less complicated and more user-friendly.



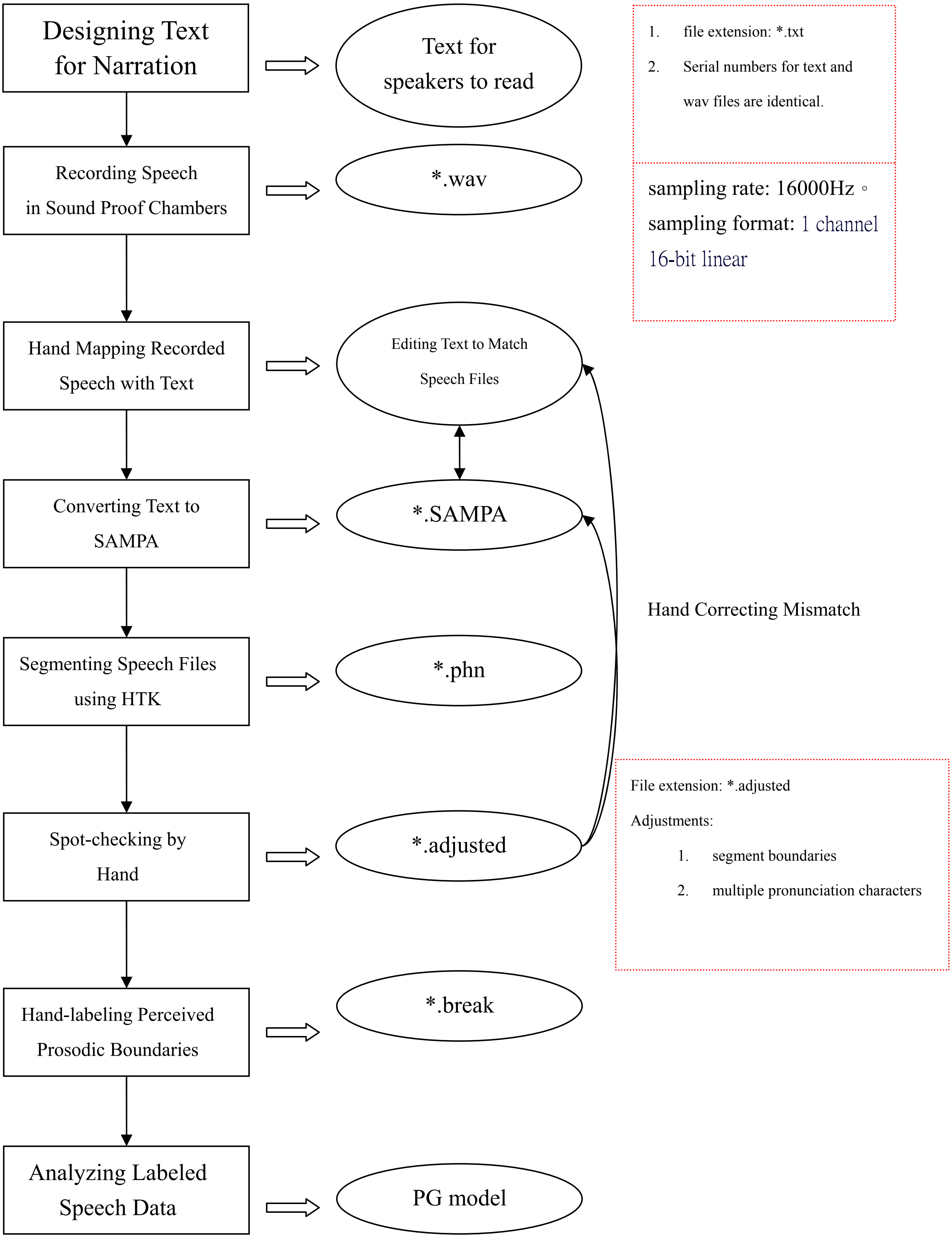
Celine - [f051p017.wav]
File(F) Edit(E) View(V) Analyze Window(W) Help(H) Debug
Wave f051p017.wav
EpochPulse fromManipulati
F0 (generated by Praat)
Formant (generated by Praa
RMS (generated by Praat)
Spectrogram (generated by
Label f051p017.adjust
Label f051p017.break
Duration (generated from la
B2 B2 B3 B3 B2 B2 B3 B3 B2 B2 B2 B2 B2 B4
NUM
Acoustic Analysis
Label / Tagging
Re-synthesizing Acoustic Signal

The speech community knows all too well how labor intensive data collection and platform design are in speech research. Dr. Chiu-yu Tseng’s working experiences on fluent speech prosody proved that research orientation and the design of data collection are the two sides of the same coin, each depending on the other. If one only collects short and isolated simple sentences and works on sentence intonations only, one will not be able to find prosody features manifested in spoken discourses through fluent speech. She also noted that in spite of a relatively large amount of existing speech databases, corpora of fluent speech remain lacking in general. A firm believer of resource sharing through public channels, especially when the original basic research is funded by government sources such as Academia Sinica, she believes the release of Sinica CORSPRO is a concrete and first step towards meeting those two principles. Sinica COSPRO and Toolkit will be released through L Labs Inc. (The Language Labs, Inc.), Taipei, Taiwan. For more information please consult URL http://www.aclclp.org.tw/use_mat_c.php#cospro .

## ● Flow chart of speech data processing and annotation

### Read speech

Task flow    output files and file names    footnotes


Designing Text for Narration
Text for speakers to read
Recording Speech in Sound Proof Chambers
*.wav
Hand Mapping Recorded Speech with Text
Editing Text to Match Speech Files
Converting Text to SAMPA
*.SAMPA
Segmenting Speech Files using HTK
*.phn
Spot-checking by Hand
*.adjusted
Hand-labeling Perceived Prosodic Boundaries
*.break
Analyzing Labeled Speech Data
PG model
Hand Correcting Mismatch

1. file extension: *.txt
2. Serial numbers for text and wav files are identical.

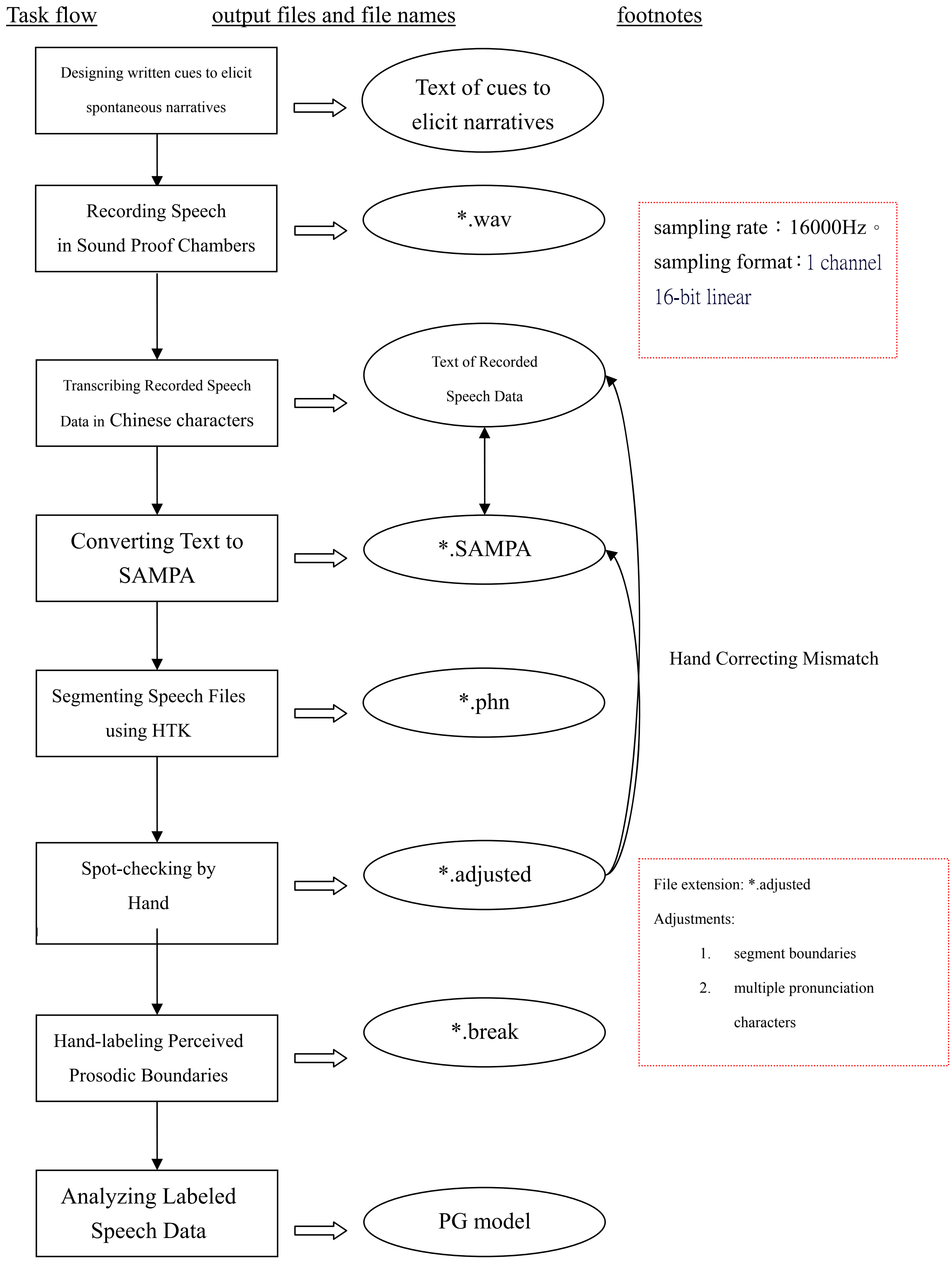
sampling rate: 16000Hz。
sampling format: 1 channel 16-bit linear

File extension: *.adjusted

Adjustments:

1. segment boundaries
2. multiple pronunciation characters

## Spontaneous Speech


Task flow
output files and file names
footnotes
Designing written cues to elicit spontaneous narratives
Text of cues to elicit narratives
Recording Speech in Sound Proof Chambers
*.wav
Transcribing Recorded Speech Data in Chinese characters
Text of Recorded Speech Data
Converting Text to SAMPA
*.SAMPA
Segmenting Speech Files using HTK
*.phn
Hand Correcting Mismatch
Spot-checking by Hand
*.adjusted
Hand-labeling Perceived Prosodic Boundaries
*.break
Analyzing Labeled Speech Data
PG model

sampling rate：16000Hz。
sampling format：1 channel 16-bit linear

File extension: *.adjusted

Adjustments:

1. segment boundaries
2. multiple pronunciation characters

## ● Processing and Labeling of Recorded Speech Files

### • Labeling System

### ◆ Machine Readable Transcription System for Chinese Dialects Spoken in Taiwan──SAMPA-T

- ✧ To process the collected speech data into annotated files for further analysis, it was necessary to transcribe the segments into machine readable phonetic transcriptions comparable to IPA (The International Phonetic Alphabets http://www2.arts.gla.ac.uk/IPA/ipa.html ) transcriptions. At the time when we began our data processing work, existing ASCII versions of phonetic transcription systems such as OGI and SAMPA (Speech Assessment Method and Phonetic Alphabets http://www.phon.ucl.ac.uk/home/sampa/home.htm ) were aimed at languages that were distinctly different from Chinese phonetically. OGI was designed exclusively for American English and SAMPA for European languages. Both proved to be insufficient to accommodate syllable-based tone languages such as Mandarin Chinese and other Chinese dialects. We therefore designed an ASCII encoding for broad phonetic transcription system for Mandarin Chinese spoken in Taiwan and also two other major dialects on the island, namely, Southern Min (Taiwanese) and Hakka, and gave the name SAMPA-T (for more information, see Tseng and Chou, 1999) and submitted the systems to SAMPA. A similar system called SAMPA-C was later designed by Chinese colleagues for Putonghua; the system was submitted to SAMPA as well.
- ✧ Though based on the principles of the IPA, SAMPA-T includes two levels of transcription, i.e., segmental and prosodic. The consequence is a more elaborate system than the IPA or its other equivalents. It is also more language dependent than IPA. These features are reflected in the design.

**Mapping between SAMPA-T and the Chinese Phonetic Alphabet (guo2yu3 zhu4in1fu2hao4), Pinyin and IPA is shown below.**

**Consonants:**

| **Guo2yu3 zhu4in1fu2 hao4** | Pin-Yin | IPA | SAMPA-T | EXAMPLES | | |
|---|---|---|---|---|---|---|
| | | | | character | syllable | meaning |
| ㄅ | b | p | b | 爆 | bau4 | to explode |
| | | b | B | 肉(T) | Ba?4 | meat |
| ㄆ | p | $p^h$ | p | 泡 | pau4 | bubble |
| ㄉ | d | t | d | 倒 | dau4 | to pour |
| ㄊ | t | $t^h$ | t | 套 | tau4 | cover over |
| ㄍ | g | k | g | 告 | gau4 | to tell |
| | | ḍ | G | 阮(T) | Gun2 | we |
| ㄎ | k | $k^h$ | k | 銬 | kau4 | handcuff |
| ㄈ | f | f | f | 斧 | fu3 | ax |
| ㄏ | h | x | h | 虎 | hu3 | tiger |
| ㄪ | | v | v | 衛(H) | vi5 | to guard |
| ㄙ | s | sɒ | s | 速 | su4 | quick |
| ㄕ | sh | ᴋ | s` | 樹 | s`u4 | tree |
| ㄒ | x | ə | sj | 細 | s\i4 | thin, fine |
| ㄖ | r | ə | Z` | 入 | z`u4 | to enter |
| | | dz,dì | DZ | 子(T) | DZi2 | small pieces |
| ㄐ | j | tə | dj | 雞 | dz\i1 | chicken |
| ㄑ | q | tə$^h$ | tj | 七 | ts\i1 | seven |
| ㄗ | z | ts | dz | 租 | tzu1 | to rent |
| ㄓ | zh | tᴋ | dz` | 豬 | tz`u1 | pig |
| ㄘ | c | $ts^h$ | ts | 粗 | tsu1 | rough, big |
| ㄔ | ch | tᴋ$^h$ | ts` | 出 | ts`u1 | to exit |
| ㄇ | m | m | m | 木 | mu4 | wood |
| ㄋ | n | n | n | 怒 | nu4 | anger |
| ㄫ | ng | ú | N | 迎(T) | Nia5 | to meet |
| ㄬ | ngi | ə | J | 你(H) | Ji2 | you |
| ㄌ | l | l | l | 錄 | lu4 | to record |

Table 1. Mapping between SAMPA-T and the Chinese Phonetic Alphabet (guo2yu3 zhu4in1fu2hao4), Pinyin and IPA, where T denotes Southern Min (Taiwanese) and H Hakka.

**Vowels:**

| **Guo2yu3 zhu4in1fu2 hao4** | Pin-Yin | IPA | SAMPA-T | EXAMPLES | | |
|---|---|---|---|---|---|---|
| | | | | character | syllable | meaning |
| ㄧ | i | i | i | 椅 | i3 | chair |
| ㄨ | u | u | u | 五 | u3 | five |
| ㄩ | yu | y | y | 雨 | y3 | rain |
| ㄚ | a | a | a | 啞 | ia3 | mute |
| | | t | E | 扁 | biEn3 | flat |
| ㄛ | o | o | o | 我 | uo3 | I |
| | | ? | ? | 懂 | doN3 | to know |
| ㄝ | e | e | e | 也 | ie3 | also, too |
| ㄜ | e | z | @ | 餓 | @4 | hungry |
| ㄦ | er | z ə | @` | 二 | @`4 | two |
| 帀 | i | ɡ | U | 絲 | sU1 | silk |
| 帀 | i | | U` | 詩 | s`U1 | poem |
| ㄞ | ai | ai | ai | 百 | bai3 | hundred |
| ㄟ | ei | ei | ei | 北 | bei3 | north |
| ㄠ | ao | o u | au | 咬 | iau3 | to bite |
| ㄡ | ou | ou | ou | 有 | iou3 | to have |
| ㄢ | an | an | an | 鹽 | ian_2 | salt |
| ㄤ | ang | o ú | aN | 羊 | iaN_2 | goat, sheep |
| ㄣ | en | z n | @n | 分 | f@n_1 | minute |
| ㄥ | eng | z ú | @N | 風 | f@N_1 | wind |
| | | nə ḍ | n^ | 黃(T) | n^5 | yellow |
| | | mɪ | m^ | 梅(T) | m^5 | plum |
| | | ú | N | 洋(T) | iaN7 | ocean |
| | | I | ~ | 贏(T) | ia~7 | to win |
| | | Ⅱ | } | 盒(T) | ap}8 | box |
| | | Ⅲ | | 力(T) | lat}8 | force |
| | | | ? | 蠟(T) | la?8 | wax |
| | | | | 六(T) | lak}8 | six |
| | | | | | | |

Table 2. Mapping between SAMPA-T and the Chinese Phonetic Alphabet (guo2yu3 zhu4in1fu2hao4), Pinyin and IPA, where T denotes Southern Min (Taiwanese), H Hakka, I nasalization of the preceding vowel, Ⅱ unreleased preceding stop consonants; and Ⅲ glottal stop.

**Tones: Numerals 1, 2, 3 and 4 denote Mandarin Chinese first (high level), second (mid rising), third (falling rising) and fourth (falling) tones, respectively. Each syllable is followed by a numeral without spacing to indicate its tone.**

- Automatic segmentation of speech files using the HTK Toolkit, followed by human spot-check
  - ✧ The HTK Toolkit was used to automatically segment speech files into phones, and labeled with SAMPA-T symbols.
  - ✧ After speech files were segmented, trained transcribers the spot checked the files to (1.) align segment boundaries and (2.) correct segmentation errors.

- Manual Labeling of Prosodic Properties—Perceived boundary breaks and prosodic units across speech flow
  - ✧ The first and most important feature of our prosody framework as reflected in our labeling system is the concept of prosodic units and boundaries in speech flow. We set out to find out how we listen, what we listen and what we hear and subsequently identify, consistently across listeners, how as speech flows the speaker would stop and then go on, where the speaker would finally make a complete stop during narratives, pause, and then how the speaker would start speaking again. While all of the above was going on, we noted that there existed a quality of lucidity and continuity as the narrative or discourse flowed. We then decided to treat such a rather big unit as a prosody unit and work from there. The result is a semantic rather than syntactic unit that spans across several (and sometimes quite a few) phrases which we identified as speech paragraphs. From the evidences we found over time, we postulated a prosody unit for fluent speech that rides above sentence intonation, and called it the Prosodic Phrase Group (PG). This top-down perspective separates us from commonly adopted approaches to speech signals or phonetic research. Note that both speech and phonetic approaches have been used to taking small fragments of speech such as segments, syllables, words and no more than phrases at a time, removing them from the speech flow, often stopping short at "local" descriptions without addressing more "global" characteristics and the relationship these local characteristics must bear with each other in fluent speech. In other words, a bottom-up perspective has been the dominating perspective. In our case, we adopted a top-down perspective instead. Note that the bottom-up approach often inadvertently treats units under investigation as unrelated entities while any top-down approach must provide evidences of governing constraints from higher nodes and hence associates the lower units as

sister constituents. As a result, a bottom-up approach could opt for either linear or hierarchical correlations among the units treated while the top-down approach would most logically only lead to a hierarchical relationship while both approaches share the same linear output form. In our case, the identified unit turns out to be multiple-phrase speech paragraphs. Hence, the most significant feature of our fluent speech prosody framework is to systematically establish cross-phrase prosodic relationships rather than dwelling on sentence or phrase intonation patterns. We noted from very early on that these speech paragraphs could consistently be distinguished by where they began and ended, and were most notably marked by perceived patterns of boundary pauses and breaks. We then proceeded to prove that these perceived breaks possess properties consistent across listeners (transcribers) and should be useful to pursuits in speech recognition as well.

- The second important feature is how our framework treats the pauses and boundary breaks across phrases within and across the identified speech paragraphs and establish their relationship to fluent speech prosody. This perspective led us to study where pauses were made during speech flow and where a speaker changed breath while speaking. We must breathe as we speak, which means speech production must interact with physiological constrains of breathing cycles. Furthermore, we also noticed from our speech data that a breathing cycle was not necessarily the ultimate speaking unit because many speech paragraphs did not end after the speaker took a complete breath, but went on afterwards. This implied that the ultimate constraint and unit of speech planning must also be cognitively based instead of physiologically based only. In summary, our prosody labeling included phonetic, physiological as well as cognitive considerations, and hence set our annotation system somewhat apart from other systems. Our subsequent prosody framework was based on evidences found on how these aspects interacted, and how these aspects were reflected through fluent speech prosody.
- In designing the labeling system, we adopted the ToBI (Tone and Break Indices http://www.ling.ohio-state.edu/~tobi ) structure where different layers of prosodic labels tag one kind of prosodic information only, but modified the tags to reflect breathing and multiple-phrase speech paragraphs in narratives or spoken discourse. Below is the break index and descriptions.

**Break Index**

|  | Definition | Characteristics |
|---|---|---|
| B0 | reduced syllabic boundary | Syllable truncation often occurred in fast or informal fluent speech. |
| B1 | normal syllabic boundary | Usually with no identifiable pauses, but more of a psycholinguistic unit for native speakers. |
| B2 | prosodic word boundary | Perceived as a boundary where a slight tone of voice change usually follows. |
| B3 | prosodic phrase boundary | A clearly perceived pause. |
| B4 | breath group boundary | Perceived end of exhale cycle followed by inhaling to begin another breathing cycle. It could be where a speech paragraph ends where trailing occurs with final lengthening coupled with weakening of speech sounds. But the speaker may still go on by breathing but not ending the speech paragraph. |
| B5 | prosodic group boundary | A complete speech paragraph ends by final lengthening coupled with weakening of speech sounds. The speaker makes a complete stop, take a new breath, and begin a new speech paragraph. |

## ● Prosody Analysis Platform--COSPRO Toolkit

The COSPRO Toolkit is a platform that integrated commonly accessible speech analysis software Adobe Audition, Praat and Speech Viewer into one common platform. The main goal for developing COSPRO Toolkit is to re-synthesize speech signals in prosodic units by extracting acoustic parameters and perceived boundary breaks. Hence, re-synthesis is the most important function in COSPRO Toolkit. The platform consists of three major functions: (1) performing acoustic analysis, (2) labeling continuous fluent speech and (3) re-synthesizing speech signals.

To perform acoustic analysis, parameters in the COSPRO Toolkit are generated by Praat. The parameters include fundamental frequency, intensity, formant frequency, and spectrogram.

To label continuous fluent speech, the COSPRO Toolkit is extremely user friendly. It maintains characteristics of Speech Viewer, but adds an object tray so that editing, playing audio output and labeling are all done by clicking. The platform is also capable of accommodating additional user-defined labels/tags in addition to existing labels.

To re-synthesize speech output, duration adjustment is performed on prosodic units of different sizes as defined by break labeling, and F0 contour pattern is performed by editing PitchTier files.

In summary, the COSPRO Toolkit is a user-friendly speech analysis software and interface in acoustic parameters and labeling functions to process and annotate fluent speech data. Furthermore, it does not require complex commands or lengthy steps to re-synthesize fluent speech.

## ● File Name Configurations

### ✧ File Name Configurations:

◆ **COSPRO serial number_speaker by gender serial number_speech data type serial number**

✧ There are 9 sets of speech corpus in COSPRO, serially named from COSPRO 01 to COSPRO 09. Speakers are represented by gender and serial number. Type of speech data is abbreviated, followed also by serial number. However, speech data type is optional. For example, COSPRO 05_M051_prg 001 denotes COSPRO speech corpus number 5 Lexically-Balanced Speech Corpus, male speaker 051, speech paragraph

001. However, speech data type is marked to reflect design purpose and is optional.

- ✧ Speech corpora and serial numbers

| Speech Corpus | Serial Number |
|---|---|
| Phonetically-Balanced Speech Database | COSPRO 01 |
| Multiple-Speaker Speech Corpus | COSPRO 02 |
| Intonation-Balanced Speech Corpus | COSPRO 03 |
| Stress-pattern Balanced Speech Corpus | COSPRO 04 |
| Lexically-Balanced Speech Corpus | COSPRO 05 |
| Focus-Balanced Prosody Group Speech Corpus | COSPRO 06 |
| Text-Type/Speaking- Style Varied Speech Corpus | COSPRO 07 |
| Prosody Balanced Monosyllable Corpus | COSPRO 08 |
| Comparable Spontaneous/Read Speech Corpus | COSPRO 09 |

- ✧ Speaker serial numbers (F：female speaker；M：male speaker）

| Speech Corpus | Sex of speaker | Number of speakers | Speaker serial number |
|---|---|---|---|
| Phonetically-Balanced Speech Corpus | M | 3 | M01~M03 |
| | F | 3 | F01~F03 |
| Multiple-Speaker Speech Corpus | M | 40 | M001~M040 |
| | F | 50 | F001~F050 |
| Intonation-Balanced Speech Corpus | M | 2 | M002~M003 |
| | F | 3 | F001~F004 |
| Stress-pattern Balanced Speech Corpus | M | 1 | SM01 |
| | F | 1 | SF01 |
| Lexically-Balanced Speech Corpus | M | 1 | M051 |
| | F | 1 | F051 |
| Focus-Balanced Prosody Group Speech Corpus | M | 1 | M052 |
| | F | 1 | F052 |
| Multiple Text-Type/Speaking- Style Varied Speech Corpus | M | 1 | M053 |
| | F | 1 | F053 |
| Prosody Balanced Monosyllable | M | 1 | M054 |

|  |  |  |  |
|---|---|---|---|
| Corpus | F | 1 | F054 |
| Comparable Spontaneous/Read Speech Corpus | M | 1 | M053 |
|  | F | 1 | F053 |

✧ Abbreviation and serial number by types of speech data

| Type of speech data | Abbreviation and serial number |
|---|---|
| Speech paragraph | prg001 |
| Spontaneous speech | sptn001 |
| Spontaneous vs. read speech | sptn001_read |
| Utterances or phrases | phr001 |
| Declarative sentences | phr001_d |
| Interrogative sentences | phr001_i |
| Exclamatory sentences | phr001_e |
| Carrier sentences | phr001_c |
| Word salad with punctuation | phr001_wd/p |
| Word salad without punctuation | phr001_wd/np |

# The Speech Corpora

## COSPRO 01: Phonetically-Balanced Speech Corpus

| | |
|---|---|
| *Item Name:* | Phonetically-Balanced Speech Database |
| *Author:* | Chiu-yu Tseng |
| *Corpus No.:* | COSPRO 01 |
| *ISBN:* | |
| *Data Type:* | speech |
| *Speaker ID:* | F01-03 ; M01-03 |
| *Sample Rate:* | 16000 Hz |
| *Sampling Format:* | 1 channel 16-bit linear |
| *Data Source(s):* | microphone |
| *Project(s):* | Collaborating Researches on Chinese Information Processing-Subproject on Mandarin Chinese Speech Database (1994.7-1999.7) |
| *Application(s):* | speech recognition, speech synthesis, pronunciation modeling |
| *Language(s):* | Taiwan Mandarin |
| *Language ID(s):* | CHN |
| *Size of digitized data:* | 1.99GB |
| *Year recorded:* | 1994 |
| *Recording duration:* | 18:38 |
| *Recording equipment:* | 1. SONY MZ-R2 MD tape recording<br>2. beyerdynamic M69N(C) mic<br>3. TDK MD tapes |
| *Online demonstration:* | yes |

### Introduction

The database was designed and collected in 1994. The goals were (1.) to obtain large amount of speech data from relatively small number individual speakers in order to study speaker related features in detail, and (2.) to obtain possible prosody features for application in speech recognition. The text part has served as the basis for the MAT (Mandarin across Taiwan) speech database project subsequently, as well as much of our own later works at the lab. The aim of the design was to obtain both phonetic and prosody information usually not available in canonical-form lexical words or phrases under 10 syllables. Text of COSPRO_01 consists of two parts: 599 discourses and a list of 1455 words. To compose text of discourses for narration, we first selected by software the most frequently used 27,000 or so lexical items (words) from CKIP's Academia Sinica Balanced Corpus of Modern Chinese http://www.sinica.edu.tw/SinicaCorpus/, then hand tailored them to compose discourses in text. Five factors were controlled: (1.) we included all possible syllables in Mandarin, less than 1300 in total, (2.) we chose the most frequently used 2- to 4-syllable lexical words, (3.) we included all possible segmental combinations and concatenations, and (4.) we incorporated all possible tonal combinations and concatenations, and (5.) we controlled discourse length at 1 to 180 characters per discourse. A total of 599 discourses were composed. We also created a word list of 1455 most frequently used words from the 27,000-word set. Each set of 1455 words plus 599 discourses required about 7 hours reading time per speaker.

**Data**

Speakers were instructed to read into the microphone in normal speaking rate at sound proof chambers at the Phonetics Lab, Institute of Linguistics, Academia Sinica. Speech data from 3 male and 3 female speakers were collected. Each male-female pair was defined by age to denote three generations of speakers. The age groups were under 25, around 35, and over 60 years of age, respectively. One of the male speakers was a retired radio announcer; the others were untrained native speakers.

The major finding from this set of speech data is the irregular F0 contour patterns at the phrase level which could not be described by definition of intonation. That is, if fluent speech prosody is analyzed phrase by phrase, then the three intonation patterns, namely, declarative, interrogative and exclamatory are hardly sufficient to describe or account for the F0 contour patterns in the corpus. From this first set of speech data, it is the multiple-phrase speech paragraphs, consistently identified across listeners that lead to a top-down perspective to process fluent speech prosody in terms of perceived units and boundaries, and the pauses between. Furthermore, how these units and boundaries are related to each other via prosodic properties in connected speech, how individual phrase intonation cannot be taken as unrelated prosody units in running speech must be addressed. This database was the first corpus of Mandarin Chinese fluent speech. It is useful for prosody as well as discourse investigations.

# COSPRO 02: Multiple-Speaker Speech Corpus

| | |
|---|---|
| *Item Name:* | Multiple-Speaker Speech Corpus |
| *Author:* | Chiu-yu Tseng |
| *Corpus No.:* | COSPRO 02 |
| *ISBN:* | |
| *Data Type:* | speech |
| *Speaker ID:* | F001-050 ; M001-040 |
| *Sample Rate:* | 16000 Hz |
| *Sampling Format:* | 1 channel 16-bit linear |
| *Data Source(s):* | microphone |
| *Project(s):* | Collaborating Researches on Chinese Information Processing-Subproject on Mandarin Chinese Speech Database (1994.7-1999.7) |
| *Application(s):* | speech recognition, speaker verification |
| *Language(s):* | Taiwan Mandarin |
| *Language ID(s):* | CHN |
| *Size of digitized data:* | 2.08GB |
| *Year recorded:* | 1996 |
| *Recording duration:* | 19:29 |
| *Recording equipment:* | 1. SONY MZ-R2 MD tape recording<br>2. beyerdynamic M69N(C) mic<br>3. TDK MD tapes |
| *Online demonstration:* | yes |

## Introduction

This database was designed and collected in 1996. The purpose was to collect relatively smaller amount of speech data (in comparison with COSPRO 01) from larger numbers of speakers in order to study segmental features and speaker variation for possible application in speech recognition research. The text was designed mainly to obtain segmental information. Each set of narrative included 83 words, 100 short sentences and 5*10 paragraphs. The 83 words were selected from the 1455 words in the Phonetically-Balanced Text (COSPRO 01); the 100 sentences from the

Intonation-Balanced Text (COSPRO 03), and the 50 paragraphs again from the Phonetically-Balanced Text (COSPRO 01).

### Data

Speakers were instructed to read into the microphone in normal speaking rate at sound proof chambers at the Phonetics Lab, Institute of Linguistics, Academia Sinica. Speech data a total of 90 speakers, 40 males and 50 females, were collected.

This speech database is useful for speaker variation both in segmental information as well as prosody phenomena of relatively short utterances.

## COSPRO 03: Intonation-Balanced Speech Corpus

| | |
|---|---|
| *Item Name:* | Intonation-Balanced Speech Corpus |
| *Author:* | Chiu-yu Tseng |
| *Corpus No.:* | COSPRO 03 |
| *ISBN:* | |
| *Data Type:* | speech |
| *Speaker ID:* | F001-004 ; M002-003 |
| *Sample Rate:* | 16000 Hz |
| *Sampling Format:* | 1 channel 16-bit linear |
| *Data Source(s):* | microphone |
| *Project(s):* | Collaborating Researches on Chinese Information Processing-Subproject on Mandarin Chinese Speech Database (1994.7-1999.7) |
| *Application(s):* | speech recognition |
| *Language(s):* | Taiwan Mandarin |
| *Language ID(s):* | CHN |
| *Size of digitized data:* | 2.38GB |
| *Year recorded:* | 1997 |

| | |
|---|---|
| *Recording duration:* | 31:19 |
| *Recording equipment:* | 1. SONY MZ-R2 MD tape recorder<br>2. AKG C410 headset mic<br>3. TDK MD tapes |
| *Online demonstration:* | yes |

## Introduction

The database was designed and collected in1997. The purpose of this database was to examine the role of intonation with respect to prosody grouping in fluent Mandarin Chinese. During our analysis of the Phonetically-Balanced Speech Database, we noticed that on the one hand, the 599 phonetically and tonally balanced paragraphs proved to be insufficient for prosody investigation in the corpus sense for lack of a somewhat balanced distribution among phrase/sentence types, namely, declarative, interrogative and exclamatory. We also noticed, on the other hand, that the speech data we collected showed a clear grouping of utterances into perceptually identifiable but larger-than-phrase/sentence prosody units in speech flow. We termed the phenomenon Prosody Group and subsequently designed another text to obtain phrase-type as well as prosody-adverbial/particle balance. That is, on the basis of three sentence types, i.e., declarative, interrogative and exclamatory, we included all possible adverbials and particles within each type to exhaust possible occurrences and to further investigate variations. Four factors were controlled, namely, (1.) utterance type, (2.), particles and adverbials, (3.) distribution of utterance type and (4.) utterance length. This speech database was our first database to concentrate on the grouping effect in Mandarin speech and its prosodic characteristics. The entire text was hand tailored to make the paragraphs as close to spoken form as possible, removing and editing occurrence of literary expressions.

A total of 1654 phrase groups were generated, including 805 declarative sentences, 546 interrogative sentences and 303 exclamatory sentences, respectively. These phrase groups ranged from 5 to 134 characters each in length. Since declaratives do not involve special particles and adverbials, the selection was based on lexical balance again from the CKIP database. Note that the utterances in this database were not as long as the Phonetically-Balanced Speech Corpus (COSPRO 01), but they were still longer utterances nonetheless.

### Data

Speakers were instructed to read into the microphone in normal speaking rate at sound proof chambers at the Phonetics Lab, Institute of Linguistics, Academia Sinica. Speech data from 3 male and 4 female speakers were collected. All of the speakers were untrained native speakers. But, there are the data of five speakers (2 males and 3 females) released in the package.

This database is useful for intonation analyses directed toward sentence types. One of the features is that instead of simple short sentences often found in phonetic studies, we also used relatively longer and more complicated sentences to hopefully render more prosodic information at the same time. This database is useful both for speech synthesis and speech recognition.

## COSPRO 04: Stress-pattern Balanced Speech Corpus

| | |
|---|---|
| *Item Name:* | Stress-pattern Balanced Speech Corpus |
| *Author:* | Chiu-yu Tseng |
| *Corpus No.:* | COSPRO 04 |
| *ISBN:* | |
| *Data Type:* | speech |
| *Speaker ID:* | sf01 ; sm01 |
| *Sample Rate:* | 44100 Hz |
| *Sampling Format:* | 1 channel 16-bit linear |
| *Data Source(s):* | microphone |
| *Project(s):* | Knowledge Representation and Language Engineering for Mandarin Chinese --- Man-machine Voice Interface Environment and Its Tools (1997.7—2002.6) |
| *Application(s):* | speech recognition |
| *Language(s):* | Taiwan Mandarin |
| *Language ID(s):* | CHN |
| *Size of digitized data:* | 243MB |
| *Year recorded:* | December 2000 |
| *Recording duration:* | 0:48 |

*Recording equipment:* 1.dbs386 Tube Pre digital amplifier
2.Creamw@re Pulsar recording sound card
3. AKG C410 headset mic

*Online demonstration:* yes

## Introduction

The database was designed and collected in 2000.This database was designed to obtain stress information for our studies of focus and prominence in speech flow and further understand the relationship between lexical stresses from utterance focus, we designed a stress balanced database and subsequently tagging systems for prominence as well . Three factors were controlled for stress balance, namely, (1.) stress type, (2.) cross-listener perceptual consistency and (3.) duration of lexical items. We chose from the above two texts lexical words ranging from 2 to 7 characters and balanced the stress distribution among the lexical items. Since the distribution of lexical words from the first two texts was insufficient to cover stress distribution, we also added more lexical word from the CKIP database and drew examples from the media for update. A total of 161 phrase groups were generated, ranging from 9 to 66 characters in length. This database is useful towards speech synthesis.

## Data

Speakers were instructed to read into the microphone in normal speaking rate at sound proof chambers at the Phonetics Lab, Institute of Linguistics, Academia Sinica. Speech data from 1 male and 1 female untrained native speaker were collected.

# COSPRO 05: Lexically-Balanced Speech Corpus

*Item Name:* Lexically-Balanced Speech Corpus

*Author:* Chiu-yu Tseng

*Corpus No.:* COSPRO 05

*ISBN:*

*Data Type:* speech

| | |
|---|---|
| *Speaker ID:* | F051 ; M051 |
| *Sample Rate:* | 16000 Hz |
| *Sampling Format:* | 1 channel 16-bit linear |
| *Data Source(s):* | microphone |
| *Project(s):* | New Directions for Mandarin Speech Synthesis : From Prosodic Organization to More Natural Output (January 2003—December 2005) |
| *Application(s):* | speech recognition, speech synthesis |
| *Language(s):* | Taiwan Mandarin |
| *Language ID(s):* | CHN |
| *Size of digitized data:* | 574MB |
| *Year recorded:* | December 2002 |
| *Recording duration:* | 35:50 |
| *Recording equipment:* | On location at the radio station |
| *Online demonstration:* | yes |

## Introduction

The text for recording was designed from 2001 to 2002 to include frequently used lexical items used in Taiwan and China; the speech database was collected in 2002. Although Mandarin Chinese is the official spoken language for both China and Taiwan, it is common knowledge among native speakers that many lexical items differ. By lexical balance here, we mean coverage and distribution of lexical words used in Taiwan Mandarin and Beijing Mandarin (or Putonghua). This aspect is essential for development in Mandarin speech technology to obtain some systematic knowledge of the lexical difference as well as pronunciation variation. In order to achieve lexical balance for both Taiwan Mandarin and Putonghua, we obtained text of recording materials from Tsinghua University at Beijing, and constructed text that covered most frequently used lexical items for both.

The lexically-balanced speech database on our side included 217 phonetically balanced (9-20 characters) sentences, 26 paragraphs (85-982 characters) and 1000 relatively short sentences (16-25 characters). The 217 sentences were selected from the MAT (http://rocling.iis.sinica.edu.tw/ROCLING/MAT/MAT-160-brief.html ) text, originally constructed at our lab. The 26 paragraphs came from two sources. We hand compiled 23 paragraphs from our previous 599 phonetically balanced paragraphs and added 3 paragraphs constructed by Tsinghua University at Beijing. The 1000-sentence set was materials from Tsinghua University at Beijing. Both COSPRO 01 and COSPRO 05 are useful for speech synthesis and TTS, especially with the concept of using prosodic units rather than unrelated monosyllables of as the basic synthesis unit. COSPRO 01 and COSPRO 05 can also be used together to study different speaking rates.

### Data

Speakers were instructed to read into the microphone in normal speaking rate at sound proof chambers at the Phonetics Lab, Institute of Linguistics, Academia Sinica. Speech data from 1 male and 1 female radio announcer were recorded. Both were under 35 years of age at the time of recording.

## COSPRO 06: Focus-Balanced Prosody Group Speech Corpus

| | |
|---|---|
| *Item Name:* | Focus-Balanced Prosody Group Speech Corpus |
| *Author:* | Chiu-yu Tseng |
| *Corpus No.:* | COSPRO 06 |
| *ISBN:* | |
| *Data Type:* | speech |
| *Speaker ID:* | F052 ; M052 |
| *Sample Rate:* | 16000 Hz |
| *Sampling Format:* | 1 channel 16-bit linear |
| *Data Source(s):* | microphone |
| *Project(s):* | New Directions for Mandarin Speech Synthesis : From Prosodic Organization to More Natural Output (January 2003—December 2005) |
| *Application(s):* | speech recognition |

| | |
|---|---|
| *Language(s):* | Taiwan Mandarin |
| *Language ID(s):* | CHN |
| *Size of digitized data:* | 759MB |
| *Year recorded:* | January 2003 |
| *Recording duration:* | 7:30 |
| *Recording equipment:* | 1. dbs386 Tube Pre digital amplifier<br>2. Creamw@re Pulsar recording sound card<br>3. AKG C410 headset mic |
| *Online demonstration:* | yes |

## Introduction

The database was designed in 2002 and recorded in2003. The purpose was to further investigate the following prosody phenomena: (1.) grouping of phrases and paragraphing in speech flow, (2.) boundaries and units involved, (3.) global planning and local specification of prosody units, and (4.) focus and prominence in speech flow, and (5.) interaction between syntax, semantics and prosody. The text was designed in 2003. Speech data was collected afterwards.

We collapsed text from the texts used from COSPRO 01 to 05, and also transcriptions from our spontaneous speech data in COSPRO 09, and compiled discourses ranging from 500 to 600 characters. Each 500-600-character discourse piece was punctuated into paragraphs ranging from 75 to 150 characters. The text for a total of 18 discourses containing a total 77 prosody groups and ranging from 347 to 712 characters. This database is useful towards speech synthesis for prominence manipulation.

### Data

Speakers were instructed to read into the microphone in normal speaking rate at sound proof chambers at the Phonetics Lab, Institute of Linguistics, Academia Sinica. Speech data from 1 male and 1 female untrained native speaker each were collected. However, instructions for reading out the text were different from those used from COSPRO 01 to COSPRO 05. Three different readings of the same text were obtained in the following order. During the first reading, subjects were asked to read out the text in normal speaking rate. Before the second reading, subjects were asked to identify and hand mark on hard copies of the text the portions they intended to emphasize in each paragraphs and subsequently read accordingly. For the third reading, subjects were given the same text with hand marked emphases, this time not by themselves by the research staff at our lab, and to read out the marked emphases accordingly.

From this database, especially from the second and third reading, we discovered that contrary to reading simple short sentences, it was very hard for subjects to maintain specified focus and prominence consistently. Subsequently, labeling results were also less consistent among trained transcribers. The results led us to believe that contextual information is as important as prosodic manifestation; prosody may not be the only vehicle to achieve focus in fluent speech.

## COSPRO 07: Text-Type/Speaking-Style Varied Speech Corpus

| | |
|---:|---|
| *Item Name:* | Text-Type/Speaking-Style Varied Speech Corpus |
| *Author:* | Chiu-yu Tseng |
| *Corpus No.:* | COSPRO 07 |
| *ISBN:* | |
| *Data Type:* | speech |
| *Speaker ID:* | F053 ; M053 |
| *Sample Rate:* | 44100 Hz |
| *Sampling Format:* | 1 channel 16-bit linear |
| *Data Source(s):* | microphone |
| *Project(s):* | New Directions for Mandarin Speech Synthesis : From Prosodic Organization to More Natural Output (January 2003—December 2005) |
| *Application(s):* | speech recognition |
| *Language(s):* | Taiwan Mandarin |
| *Language ID(s):* | CHN |
| *Size of digitized data:* | 577MB |
| *Year recorded:* | May 2003 |

| | |
|---|---|
| *Recording duration:* | 1:32 |
| *Recording equipment:* | 1. HHB Portadisc MDP500 MD tape recorder<br>2. AKG C410 headset mic |
| *Online demonstration:* | yes |

## Introduction

This database was designed and recorded in 2003. The purpose was to further test grouping effect in fluent Mandarin speech by removing syntactic and semantic information altogether. Text of random character sequences was generated in the spring of 2003. Speech data from 1 male and 1 female untrained native speaker each were collected.

Simple in-house software was used to randomly select and string characters from our previous texts used in COSPRO 01 to COSPRO 06 into nonsense lexical strings from 10 to 60 characters, nicknamed word salad at our lab. Four factors were controlled, i.e., (1.) tones, (2.) function words, (3.) word frequency and (4.) pronunciation. For tones, we controlled even distribution of the 4 Mandarin tones and 1 neutral tone for the paragraph-final syllable, and then controlled the second character backwards for even distribution of the 4 tones one more time. Consideration of tone control for the last two syllable of an utterance was to balance tone distribution as well as to capture possible disyllabic effect at the lexical word level. For function words, two controls were imposed. One was to prohibit function words to occur at utterance initial position, and another was their even distribution within the text. The latter was aimed to see whether function words would affect grouping to move forward or backward in speech flow, possibly marked by insertion of a pause. In addition, the number of characters between function words remained consistent before and after function word insertion in order to avoid possible effect to phrase grouping and/or speech rhythm. In other words, Less frequent characters were removed from the text reduce hesitation or confusion in the reading task. Characters with more than one pronunciation were also removed for the same reason. All of the text generated was subsequently hand tailored to remove any possible meaningful reading or proper names.

80 sections of word salad, 25 utterances and 2 paragraphs were generated. The 80 sections of word salad ranged from 10 to 60 characters, 40 of them were presented in character strings without punctuation; the other 40 with punctuation marks randomly assigned by software and ended in periods. The 25 utterances ranged from 17 to 83 characters; the two paragraphs 393 and 461 characters respectively. Both were samples from our earlier databases and properly punctuated. These utterances and

paragraphs served as reference of the speakers' regular reading speech for the comparison purposes. This database provides monosyllables in various contextual positions without any syntactic and/or semantic information, and is also more natural than monosyllables produced in isolation. The database should be useful for concatenative synthesis of Mandarin speech.

### Data

Speakers were instructed to read into the microphone in normal speaking rate at sound proof chambers at the Phonetics Lab, Institute of Linguistics, Academia Sinica.
Speech data from 1 male and 1 female untrained native speaker each were collected.
Both speakers were under 30 years of age at the time of recording

## COSPRO 08: Prosody Balanced Monosyllable Corpus

| | |
|---|---|
| *Item Name:* | Prosody Balanced Monosyllable Corpus |
| *Author:* | Chiu-yu Tseng |
| *Corpus No.:* | COSPRO 08 |
| *ISBN:* | |
| *Data Type:* | speech |
| *Speaker ID:* | F054 ; M054 |
| *Sample Rate:* | 16000 Hz |
| *Sampling Format:* | 1 channel 16-bit linear |
| *Data Source(s):* | microphone |
| *Project(s):* | New Directions for Mandarin Speech Synthesis : From Prosodic Organization to More Natural Output (January 2003—December 2005) |
| *Application(s):* | speech recognition; speech synthesis |
| *Language(s):* | Taiwan Mandarin |
| *Language ID(s):* | CHN |

| | |
|---|---|
| *Size of digitized data:* | 1.9GB |
| *Year recorded:* | February 2004 |
| *Recording duration:* | 16:50 |
| *Recording equipment:* | 1.HHB Portadisc MDP500 MD tape recorder<br>2. Sony ECM-77B mini mic |
| *Online demonstration:* | yes |

## Introduction

This database was designed and recorded in 2004 for speech synthesis needs. While it is feasible to use prosodic units rather than monosyllables to synthesize Mandarin Chinese speech, collecting a semi-exhaustive set of polysyllabic prosodic units is not feasible for a research lab of our size and scale. To compensate insufficient amount of prosodic units, we need speech data of monosyllables from the same speakers to fill in the gap. Furthermore, it is also important to obtain Mandarin Chinese monosyllables that would bear the most distinct fluent prosodic characteristics so that they can also serve as acoustic reference for prosodic manipulations if need be. This database is particularly useful towards unlimited TTS. Below we describe the design in more detail.

1. The database is designed for speech synthesis of Mandarin Chinese via syllable concatenation.
2. Text for reading includes two parts. Part 1. Monosyllables in carrier sentences ; Part 2. Paragraphs and intonation-balanced sentences.
3. Speakers were instructed to read or speak into the microphone in normal speaking rate at sound proof chambers at the Phonetics Lab, Institute of Linguistics, Academia Sinica. Speech data from 1 male and 1 female untrained native speaker were collected. Both speakers are graduate students, bilingual native Mandarin Chinese and Southern Min speakers. They are representative of regional accent of the educated and speaking style in Taiwan. When recording began in February 2004, the male speaker (M054) was 26 years of age and the female speaker (F054) 44. Both speakers were students at the Graduate Institute of Teaching Chinese as a Second Language, National Taiwan Normal University.

## Data

### Part 1. Monosyllables embedded in carrier sentences.

1. Existing monosyllables embedded in carrier sentences（phr0001_c~phr1409_c）.

A 30-syllable 3-phrase complex sentence that represents a Prosodic Phrase Group was constructed the carrier sentence. Speakers were instructed to read in normal speaking rate with natural focus patterns.。

Control factors:

(1) A total of 1,300 existing syllables with all possible segmental and tonal variations.

(2) Target syllable is embedded in a 30-syllable 3-phrase complex carrier sentence; the three phrases represent PG-initial, -medial and –final positions to render fluent speech prosodic information.

Examples (with the target syllable in red) are shown below:

巴ㄅㄚ是一個常見的字，一般人常把巴ㄅㄚ字掛在嘴邊，講話時動不動就會提到巴ㄅㄚ。

吧˙ㄅㄚ是一個常見的字，一般人常把吧˙ㄅㄚ字掛在嘴邊，講話時動不動就會提到吧˙ㄅㄚ。

2. Sandhi monosyllables embedded in carrier sentences

Control factors:

(1) From the set of existing 1300 monosyllables, 109 Tone 3 syllables do not have a Tone 2 counter part.

(2) Di-syllabic third-tone-sequence lexical words ＸＸ containing one of the 109 monosyllables was chosen and embedded into the carrier sentence. For example:

ＸＸ是一個常見的詞，一般人常把ＸＸ掛在嘴邊，講話時動不動就會提到ＸＸ。

## Part 2. Paragraphs and Intonation Balanced Sentences

1. Control factors:

   (1) Different levels of prosodic units (see Tseng 2005)

   (2) Principles of phonetic balance

   (3) Speaking rate

2. Text design for paragraphs

   (1) Source of text: Text from Phonetically-Balanced Speech Database (COSPRO 01), Intonation-Balanced Speech Database (COSPRO 03) and Lexically-Balanced Speech Database (COSPRO 05).

   (2) Modification principles:

      A. Based on text used in COSPRO 01, prosodic units by boundary break labeling were extracted as different levels prosodic units. From these units, we calculated missing syllables from Part 1, removed the least frequently used characters, and derived 252 missing syllables. Our plan was to derive as many such syllables as possible from the Intonation-Balanced Speech Database (COSPRO 03), then construct new text pieces for remaining syllables.

B. Text from COSPRO 05 was further hand tailored and trimmed, then subsequently re-arranged into slightly different 27 paragraphs in order to compose longer discourses for more lucid continuous reading style (p001-p027).

C. The non-overlapped portion of text between COSPRO 01 and COSPRO 05 was further re-arranged into 21 paragraphs, hand tailored for more colloquial style in order to elicit more natural and lucid reading style （p028-p048）.

D. From the 48 paragraphs derived through steps B and C, we hand picked 11 relatively short paragraphs that both speakers had rendered fluent and lucid reading, and asked the speakers to read in faster or slower speaking rate. The purpose was to produce identical speech data from the same speakers in normal vs. fast/slow speaking rates for comparative studies in overall prosodic modifications in general; and temporal allocation and syllable duration patterns in particular.

E. The same text from the Intonation-Balanced Speech Corpus (COSPRO 03) was also used.

3. Data

(1) 70 paragraphs

(2) 1654 sentences from 5 to 134 syllables（805 declarative sentences, 546 interrogative sentences and 303 exclamatory sentences）

# COSPRO 09: Comparable Spontaneous/Read Speech Corpus

| | |
|---|---|
| *Item Name:* | Comparable Spontaneous/Read Speech Corpus |
| *Author:* | Chiu-yu Tseng |
| *Corpus No.:* | COSPRO 09 |
| *ISBN:* | |
| *Data Type:* | speech |
| *Speaker ID:* | F053 ; M053 |
| *Sample Rate:* | 16000 Hz |
| *Sampling Format:* | 1 channel 16-bit linear |
| *Data Source(s):* | microphone |
| *Project(s):* | New Directions for Mandarin Speech Synthesis : From Prosodic Organization to More Natural Output (January 2003—December 2005) |
| *Application(s):* | speech recognition |
| *Language(s):* | Taiwan Mandarin |
| *Language ID(s):* | CHN |
| *Size of digitized data:* | 76MB |
| *Year recorded:* | May 2003 |
| *Recording duration:* | 0:42 |
| *Recording equipment:* | 1. HHB Portadisc MDP500 MD tape recorder<br>2. AKG C410 headset mic |
| *Online demonstration:* | yes |

## Introduction

This database was designed and recorded in the spring of 2003. The goal of this spontaneous-and-read speech data was to (1.) begin investigation of prosody of spontaneous speech, (2.) compare and derive prosodic characteristics for both read and spontaneous speaking styles, and (3.) study possible paralinguistic as well as non-linguistic effects on prosody manifestation. Our focus was also running speech. As a result, relative longer monologues were collected. So far we have collected 8 discourses.

## Data

Speakers were instructed to read into the microphone in normal speaking rate at sound proof chambers at the Phonetics Lab, Institute of Linguistics, Academia Sinica. The same untrained two speakers, one male and one female, from the Text-Type/Speaking-Style Varied Speech Corpus (COSPRO 07) recorded the present Comparable Spontaneous-and-Read Speech Corpus (COSPRO 09). Both speakers were under 30 years of age at the time of recording.

Instructions for the speakers varied most for this speech database. We devised three phases of recording to achieve our goal.

**Phase 1: Read speech of word salad.**

The first phase overlapped with data collection described in COSPRO 07. That is, speakers were asked to read text of word salad. This turned out to be a difficult job for the speakers, both for the content and for the recording processes involved. But as recording time increased, the speakers became familiar with the set-up and the sound proof chamber, and grew more experienced and relaxed for the recording sessions.

**Phase 2: Spontaneous speech.**

At this point, we were ready to collect spontaneous speech data. We set out to collect spontaneous speech on specific topics as well as free monologues. For specific topics, we chose the outbreak of SARS and related reports in the spring of 2003. Each speaker was asked sight read without sounding out a piece of text we provided for a period of 30 minutes, and was allowed to take notes of the materials during sight reading. The text consisted of headlines and front page coverage on SARS. When the 30 minutes were up, each speaker was then asked to enter the sound proof chamber with their own notes to give an oral report into the microphone of what they had just read. We made a point to use head sets microphones (AKG C410) so that the speakers were not inhibited from hand gesturing and body movements while speaking. This turned out to be much easier for untrained speakers to generate narration or monologue in length. We then offered a coffee break, and the speakers were asked to record another round of free monologues. By this time our speakers were completely relaxed and further provided narration of some treasured personal experiences, with an experimenter at the side as the loyal listener, reciprocating with eye contacts and proper body language such as nodding. This concluded the second phase of data collection, during which we obtained two kinds of spontaneous speech.

**Phase 3: Read speech of transcribed text from Phase 2.**

The third phase was days later, after we made orthographic transcription of the spontaneous speech obtained during the second phase of recording. The same speakers were asked to read out text of transcription of their own previous spontaneous speech. After the third phase, we would be able to obtain both spontaneous and read speech data of identical content for future phonetic as well as prosodic comparisons. This database is useful for speech synthesis.