Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech

Chiu-yu Tseng
Phonetics Lab, Institute of Linguistics,
Academia Sinica
128 Academia Rd. Sec. 2
Taipei, Taiwan 115
cytling@sinica.edu.tw

Yun-Ching Cheng & Chun-Hsiang Chang
Phonetics Lab, Institute of Linguistics,
Academia Sinica
128 Academia Rd. Sec. 2
Taipei, Taiwan 115

Abstract

This paper reports the content and release of The Sinica COSPRO (Mandarin Continuous Speech Prosody Corpora) and Toolkit, Academia Sinica, Taipei, Taiwan (http://www.myet.com/COSPRO). The package includes a total of about 11.99 GB recorded speech corpora (7.7 GB annotated and human spot-checked) and analysis platform developed.

The focal point of our research is its perspective, namely, how to approach and account for fluent speech prosody. The corpora are reflections of a top-down perspective emphasizing the interacting relationship among on-line speech planning and cognitive limits in addition to physiological as well as articulatory constraints during speech production, exemplified only through the associative patterns within and across phrases in fluent speech prosody. Due to the interactions among these factors, fluent speech is a mixture of both robust and slurred speech signals that can not and should not be viewed simply as concatenation of unrelated prosodic units into speech strings, be they large or small. Two major characteristics distinguish our top-down hierarchical multiple phrase framework from most of other prosody analyses. One is the units and boundaries in fluent speech perceived by listeners; the other is how to implement our findings to more practical applications. We believe that any attempt to derive or simulate fluent speech prosody must account for the above factors.

Since speech data collection began as early as 1997, some of the corpora have been reported before at O-COCOSDA (Tseng et al 2003). Annotated part of the corpora was hand labeled for perceived boundaries and units, and human spot-checked.

Through the COSPRO Toolkit we also share our research method with the community. Our Toolkit is a very friendly window-based platform that integrates commonly accessible speech analysis software such as Adobe Audition, Praat and Speech Viewer into one platform. It accepts both COSPRO tagged speech files as well as user defined tags and could also be easily incorporated into existing data driven approaches to improve prosody output. The platform consists of three major functions: (1) performing acoustic analysis, (2) labeling continuous fluent speech and (3) re-synthesizing speech signals. The most important feature of COSPRO Toolkit is the re-synthesis function. Acoustic parameters can be extracted from speech signals in prosodic units from COSPRO and subsequently manipulated independently or collectively to generate speech output. Based on a modular acoustic model we constructed [1] , users are able to manipulate F0 contours, syllable durations, intensity distribution and boundary breaks across phrases, both within or across speakers. As a result, the Toolkit allows users to change the melody, rhythm and speaking rate of a speaker, or to port the above features from one speaker to another to generate different prosody output.

We believe that Sinca COSPRO and Toolkit will be very useful to the speech community, especially to technology development. Future works include further investigating prosody related phenomena such as F0 reset and modification patterns of F0 range across phrases towards speech synthesis, as well as implementing the concepts of perceived boundaries, units and cross-phrase templates to speech recognition.

1 Introduction

Why does the research community of phonetics and speech science need corpora of continuous (or fluent) speech for prosody investigations? What is special about COSPRO and the Toolkit? The answer is very simple: instead of studying speech prosody from syllables and tones upward and stopping at isolated individual intonation patterns, we studied fluent running speech of read narratives from a top-down perspective and did not limit ourselves to phase intonations only.

This perspective led us to locate

multiple-phrase speech paragraphs on top of phase intonations in the first place. Subsequently, we also obtained the following evidences regarding the organization of fluent speech prosody that addresses the association among phrases in fluent speech: (1.) multiple-phrase melodic templates in addition to single-phrase intonations, (2.) cross-phrase syllable duration cadence and speech rhythm patterns in addition to isolated syllable durations, (3.) intensity distribution patterns at the phrase levels corresponding to higher level planning, and (4.) layered boundary break patterns above the phrase level. From these evidences, templates and patterns we were able to construct a hierarchical framework of fluent speech prosody that emphasizes on cross-phrase associations. A mathematical modular model was also constructed to generate prosody. In retrospect, if we focused on intonations of simple and short simple sentences and treated them as unrelated prosody units, we would not be able to understand fluent speech prosody. By the same logic, it is no surprise that the speech technology community could not simulate fluent speech prosody by simply concatenating intonations of short utterances without further ado whereas the further-ado is mainly about what units form prosody and how the associative relationship these units bear to each other. We believe that this top-down perspective and the framework we propose is a reflection of how we humans produce and perceive fluent speech, or put simply, how the two-way cycle of speech communication is constituted.

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. A total of approximately 11.99GB of recorded speech (7.7 GB annotated and human spot-checked) is delivered and is by far the only and largest corpora of Mandarin Chinese fluent continuous speech designed to bring out features of fluent speech prosody. Almost all of the corpora are read discourses; only 80MB are spontaneous narratives. The Toolkit is a perceptually based annotation platform catered to labeling perceived boundary breaks that would derive various levels of prosodic units in fluent speech. The software is window-based and user-friendly, integrating commonly accessible speech analysis software such as Adobe Audition, Praat and Speech Viewer into one common platform. Functions include performing acoustic analysis, labeling continuous fluent speech for prosodic characteristics and most importantly, re-synthesizing speech files to manipulate prosodic output. The rationale was simple: given that speech signals are surface output of language in communication, it is necessary to take into consideration articulatory, linguistic, physiological and cognitive constraints during on-line speech production and processing as well as the speaker's intensions and the listener's interpretations. All of these factors form the necessary and minimum infrastructure of speech communication. Without some understanding of these factors, fluent speech prosody would just be acoustic signals with all too many unpredictable variations for data-driven models o handle.

Funding resources for corpus collection and toolkit development came exclusively from Academia Sinica, mostly under the support of three Academia Sinica Interdisciplinary Theme Projects, "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).

2 COSPRO Databases

There are 9 sets of Mandarin Chinese fluent speech corpora. They are: (1.) Phonetically Balanced Speech Database (COSPRO 01, 2047.8MB, 18:38), (2.) Multiple Speaker Speech Corpus (COSPRO 02, 2141MB, 19:29), (3.) Intonation Balanced Speech Corpus (COSPRO 03, 3441MB, 31:10), (4.) Stress-pattern Balanced Speech Corpus (COSPRO 04, 244MB, 48m), (5.) Lexically-balanced Speech Corpus (COSPRO 05, 568.3MB, 35:50), (6.) Focus-balanced Prosody Group Speech Corpus (COSPRO 06, 1346MB, 7:30), (7.) Text-type/Speaking-style Varied Speech Corpus (COSPRO 07, 626.7MB, 1:32), (8.) Prosody Balanced Monosyllable Corpus (COSPRO 08, 1500MB, 15:12), and (9.) Comparable Spontaneous/Read Speech Corpus (COSPRO 09, 80MB, 42m). Each corpus was designed to bring out different prosody features involved in fluent speech. 7 out of the 9 sets of corpora were reported in Oriental COCOSDA 2003 (Tseng et al, 2003), some of them under slightly different names. For more detailed information, please consult http://www.myet.com/COSPRO

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 (*adjust), (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 fluent speech prosody framework. More comprehensive information of the fluent prosody framework and modeling is available in [1].

3 COSPRO Toolkit

The COSPRO Toolkit is a platform that integrates several pieces of commonly accessible speech analysis software such as Adobe Audition, Praat and Speech Viewer into one common platform. In addition to analyzing the acoustic properties of speech signals, the most important function the Toolkit is to re-synthesize speech signals in annotated prosodic units by extracting acoustic parameters and perceived boundary breaks. Figure 1 shows the functions of COSPRO Toolkit. The platform consists of three major functions: (1) performing acoustic analysis, (2) labeling continuous fluent speech and (3) re-synthesizing speech signals. In the following units, we will go into details about these three functions.

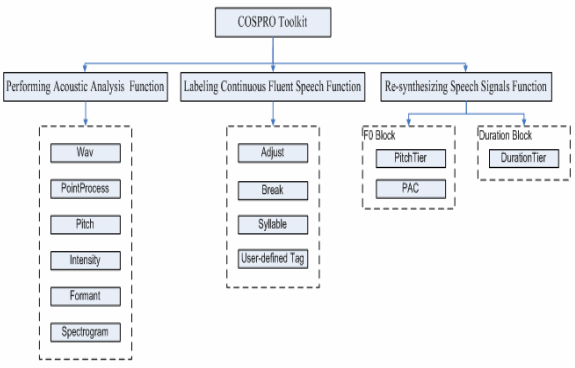

Figure1. COSPRO Toolkit functions.

3.1 Performing Acoustic Analysis

To perform acoustic analysis, acoustic parameters in the COSPRO Toolkit are generated by Praat [2]. Figure 2 shows the acoustic parameter calculation function in COSPRO Toolkit. These parameters that can be calculated include "EpochPluse", "PitchTier", "Intensity", "Formant" and "Spectrogram", each corresponding to "Pulses", "Pitch", "Intensity", "Formant" and "Spectrum" in Praat.

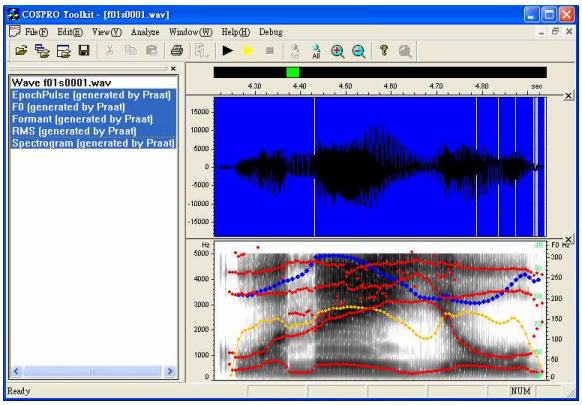

Figure2. The COSPRO Toolkit provides multiple displays of calculated acoustic parameters.

3.2 Labelling Continuous Fluent Speech

To label continuous fluent speech, the COSPRO Toolkit is designed to be extremely user-friendly. Figure 3 shows the speech labeling function in COSPRO Toolkit. It maintains characteristics of Speech Viewer [3], but adds an object tray, a function of Adobe Audition [4], so that editing, playing audio output and labeling are all done by mouse-clicking. The labeling functions not only read and edit COSPRO-defined files including adjustment, break and syllable files, but also support new user-defined labels/tags based on COSPRO provided time-code format such as prominence files. Furthermore, the Toolkit can edit traditional Chinese text in user-defined files.

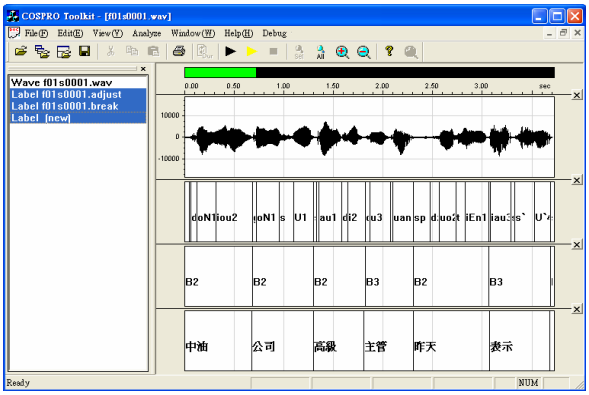

Figure3. The COSPRO Toolkit provides layered labelled results of speech signals, phonetic transcriptions in SAMPA, perceived boundary breaks and Chinese characters.

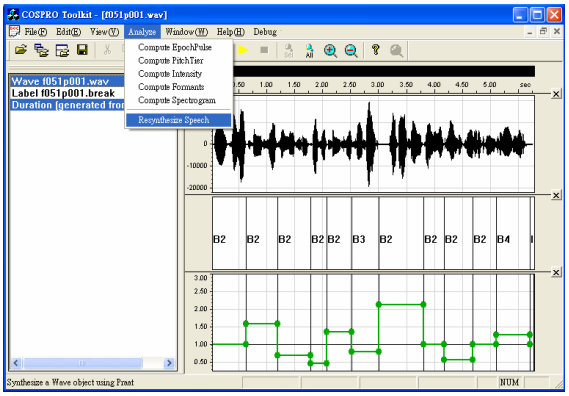

Figure4. An example of duration re-synthesis by COSPRO Toolkit.

3.3 Re-synthesizing Speech Signals

To re-synthesize speech output, COSPRO Toolkit is capable of re-synthesizing new waveform for duration and F0. Figure 4 shows the duration re-synthesis function in COSPRO Toolkit. For duration re-synthesis, we can use break files to generate a duration object that provides normalization value to users. Users can re-synthesize new waveform according to selected wave file and edited duration Tier file. For F0 re-synthesis, COSPRO Toolkit provides two methods for users, PitchTier Files by Praat and PAC files using the Fujisaki model [5]. Figure 5 shows the F0 re-synthesis function by Praat in COSPRO Toolkit. We can analyze a selected wave file to generate a Pitchtier file. Users can edit the F0 values of a PitchTier file, and re-synthesize selected wave file and edited PitchTier file into new waveform. Figure 6 shows the F0 re-synthesis function by the Fujisaki model in COSPRO Toolkit. We can see the values of Ap and Aa commands that are extracted by the Fujisaki model from Figure 6. Users can change the values of Ap and Aa commands at their will by clicking the cursor. The function can use the original wave file and edited PAC file to re-synthesize new waveform.

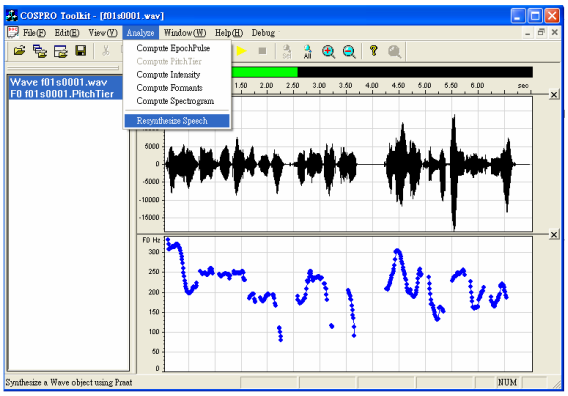

Figure 5. An example of F0 re-synthesis by Praat in COSPRO Toolkit.

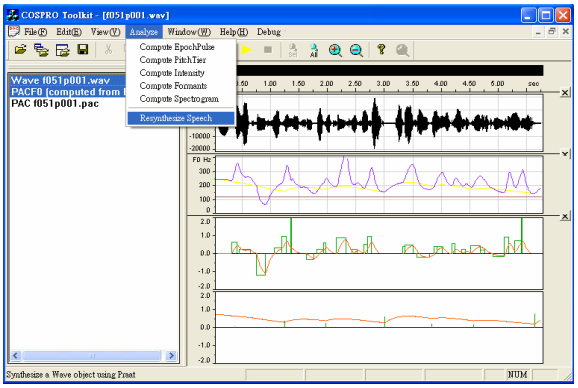

Figure6. An example of F0 re-synthesis by the Fujisaki model in COSPRO Toolkit.

4 Discussion

The major feature of the Corpora and Toolkit is its theoretical orientation, namely, the organization and implication of the prosody of continuous running speech, or for simplicity purposes hereafter, fluent speech and its effect in both speech production and perception. We believe that

any attempt to derive or simulate fluent speech prosody, especially when adopting a corpus approach, must account for the on-line speech planning and speech processing involved in narration or discourse communication. How individual phrases in fluent speech are associated through prosody is linguistically and communicatively significant, and therefore merits research attention. Analyzing fluent speech corpora from a top-down perspective on perceived units and boundaries across running speech, multiple-phrase complex sentences and/or speech paragraphs were consistently identified across listeners, instead of individual phrases. Figure 7 is a schematic representation of the hierarchical framework showing how multiple phrases are grouped into a speech paragraph. The framework was supported by evidencess from derived cross-phrase global F0 patterns, syllable cadence patterns, intensity distribution units and boundary breaks patterns in relation to higher-level prosody governing, as well as layered contributions that cumulatively formed the overall prosody output. Based on the evidences found, we postulated a multiple-phrase hierarchical prosody framework PG [1] corresponding to multiple-phrase complex sentences and/or speech paragraphs in fluent speech of narratives. A 4-modular mathematical model was also constructed to predict cross phrase F0 countour, duration allocation, intensity distritubion and boundary breaks. The framework takes in to consideration factors such as planning threshold, planning strategies, prosodic units and boundary breaks are in fact reflected in the cross-phrase melodic, rhythmic and energy distribution patterns. Furthermore, how the above-mentioned factors, together with syntactic structures, semantic interpretations and speakers' intensions, collectively contribute to the communication infrastructure of fluent speech. Due to the interactions among these factors, we view fluent speech as a mixture of both robust and slurred speech signals seamlessly associated. This view emphasizes that fluent speech can not be viewed as concatenation output of unrelated prosodic units, be they large or small. Our major research concern was how this mixture could be and should be studied.

Our research has demonstrated that a hierarchical prosody structure existed above phrases and associated them into speech paragraphs, and that is why various layers of acoustic templates exist as well. The interaction and integration of these different levels of acoustic templates must take place during on-line speech planning, and provide forecast and look-ahead references for on-line speech processing. Therefore, how fluent speech is perceived and how it should be decomposed are both critical and crucial.

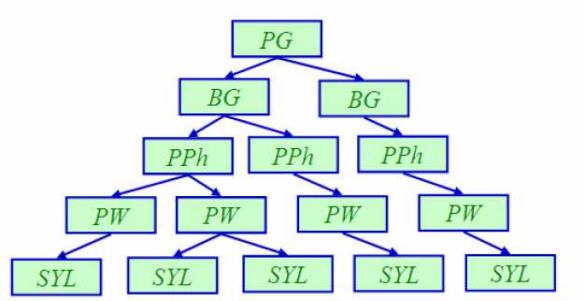

Figure7. A schematic representation of the hierarchical organization of multiple-phrase grouping (PG) and association on perceived units and boundaries.

Two major characteristics distinguish the top-down hierarchical PG framework from orientations of other prosody analyses. One is the units and boundaries in fluent speech; another is how to implement our findings to more practical applications. Both features emphasize the associative relationship between and across prosody units as speech flows fluently. On the units and boundaries in fluent speech, two acoustic domains received special attention: one is the role of phrasal intonation and another one temporal allocation across phrases. We found that individual phrasal intonations are no longer unrelated independent prosody units, but rather, sister constituents subject to higher commands from PG. On the melody and tune of fluent speech, the PG-specified positions cause phrases under grouping to modify their intonations in order to signal the beginning, continuation and finally termination of a speech paragraphs. These PG-position specifications may or may not be syntactic, often go beyond phrases or sentences, and can also be semantic or speaker-intended. On F0 contour patterns, the PG effects are delivered through cross-phrase melody cadence templates; their most significant function is to signal where a speech paragraph will end. On temporal allocation patterns, we found consistent cross-speaker and cross-phrase cadence templates of syllable duration at each and every prosodic level. In other words, 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. Both F0 and duration adjustments feature cross-phrase associations and must include levels of boundary breaks. To summarize, higher-up commands from speech paragraph cause cross-phrase F0 and duration patterns to adjust and modify. To yield

fluent speech prosody, intonation modifications alone are insufficient unless syllable durations are also modified while both modifications are in accordance with respective templates in the same way syntax constrains and generate sentences. Our findings in temporal allocations are not only significant to how Mandarin Chinese syllables should be studied in fluent speech, but also to how we can no longer stop at studying F0 contours alone and think we have done most of the work for prosody. The speaker delivers all of the above information via prosody; the listener reciprocates by picking up these prosodic cues; and the communicative cycle is complete.

Our 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 framework [1], 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.

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. We have used the model in the Toolkit, ready to be used with any synthesis program for prosody adjustments.

5. Conclusion

It is our hope that through the Sinica COSPRO and Toolkit, we share our experiences and results with research communities in phonetics, speech science and speech technology, in particular the community of technology development. On-going works include further investigating prosody related phenomena such as F0 reset and modification patterns of F0 range across phrases towards TTS, as well as implementing the top-down perspective of prosodic units in relation to boundaries, cross-phrase templates and associative relationship among prosodic units to speech recognition.

References

[1] Tseng, Chiu-yu, ShaoHuang Pin and Yeh-lin Lee, Hsin-min Wang and Yong-cheng Chen (2005) "Fluent Speech Prosody: Framework and Modeling" Speech Communication, Vol. 46:3-4 pp. 284-309

[2] http://www.fon.hum.uva.nl/praat/

[3]http://cslu.cse.ogi.edu/toolkit/docs/2.0/apps/speechview/index.html

[4] http://www.adobe.com/products/audition/

[5] Tseng, Chiu-yu and Pin, Shao-huang. **"**Modeling prosody of Mandarin Chinese fluent speech via phrase grouping**"** *Speech and Language Systems for Human Communication* (COCOSDA2004), New Delhi, India, pp.53-57, 2004