Acoustic Analysis of Dialogue Speech - Description and Verification of Spontaneous Speech and Measurement of Its Speech Rate -

Shigeyoshi Kitazawa, Satoshi Kobayashi, Hideya Ichikawa, and Motoki Sugimura
Department of Computer Science, Shizuoka University
5-1, 3 Chome, Hamamatsu, 432, Japan
e-mail: kitazawa@cs.inf.shizuoka.ac.jp

1. Introduction

Nonverbal communications play an important role in human dialogue where participants use natural speech so called spontaneous speech. That nonverbal information interchanged between people is called as "paralanguage". There are number of aspects of paralanguage. Some of them are sensational features, and difficult to realize measurements. The others, however, are measurable as concrete acoustical features. Here we described the transcription into text of the paralinguistic features such as voice loudness, tone, and speech rate. We made various measurements of validity and consistency of descriptions between different transcribers. Then we described how to measure the speech rate, one of the paralinguistic features.

2. Transcription of Dialogue Speech

In studying spontaneous speech, transcription is necessary and indispensable. Audio or video tape recording is just a material and is very insufficient for information processing. Transcription is a kind of information processing by human being, where interpersonal communication is described in terms of concepts that are important for human communication. Study of spontaneous speech is at the stage of pre-automatic recognition, where the behavior and phenomenon should be studied as they are.

There are number of ways different researchers developed to represent speech in a written forms. This is because speech has various aspects which are important for different researchers. Speech can be described according to a large number of dimensions, many of which have not been written in text forms. The approach we adopted here as a basis is transcriptions of speech in the TEI (the Text Encoding Initiative)[1].

The TEI is developed for people who works with any "text" in electronic form to process by machine and to encourage the sharing of electronic texts. The TEI is a set of encoding conventions, provides a fully general text encoding model or a framework. The guideline provides means of representing those features to facilitate processing of the text by computer programs. In particular, they specify a set of markers (or tags) which may be inserted in the electronic representation of the text, in order to mark the text structure and other textual features of interest. Without such explicit markers, many important features remain difficult to locate by mechanical means as computer programs, and thus difficult to process effectively.

The TEI began from text description has rich framework and concepts for dialogue description which we aimed at in this research. Our description framework is similar and concepts are also adopted. In particular TEI has a framework of description of nonverbal features of speech, paralinguistic features are involved using technical terms of behavioral science. Even though there are description methods of more physical parameters, for example, voice tone may be described with conventional pitch extraction method provides numerical measure of voice heights, however, such measure does not reflect human information processing, that is the paralinguistic features is measured, defined, understandable, and make sense only for human. Any machine can not recognize paralanguage. Paralanguage is in a pre-recognition stage for machine.

3. Consistency of Inter-Transcribers' Transcription[2]

Text encoded dialogue is a useful form of analysis data, however, consistency between transcribers and omission and misleading are unavoidable. We estimated these errors and inconsistencies with cross check between different transcribers on some conditions such as transcribers' skill, using graphical interface and verifying transcription with synthetic sounds based on own transcription. Their descriptions are according to the TEI encoding scheme of utterances and paralinguistic descriptions of real dialogue. We could find 91 % agreements between different transcribers of phonetic transcriptions, and 64 % of agreements of nonverbal transcriptions. These results suggest us that some acoustic parameters will help consistent description of nonverbal features.

We used two real speech data. One of the speech data was used for training of transcribers, and another was used for test. Each of the speech data has about 40 seconds' duration, and were taken from a TV program of a female interviewer and a male interviewee. Both verbal sounds and nonverbal sounds, e.g., laugh are transcribed using TEI tags and entities. All of subjects attended to meetings, to learn about TEI notation and how to transcribe the dialogue, before first session.

Consistency rates of verbal sounds marked almost the same rates through sessions. But, consistency rates of tags become higher through sessions. Training of transcribers and using GUI affect good influence for consistency rates, but, no interactions appear. Then, we made a hypothesis "Verifying own transcription by comparing synthesized sounds with original dialogue has no affect on improvement." Improvement by hearing synthesized sound is not proved except tags.

4. Qualitative Evaluation of Transcription by Synthesis of Spontaneous Dialogue

Current text to speech synthesizer produces prosodic features with a pre-determined rule from syntax. But spontaneous speech have more than predicted nonverbal features, such as tempo, pitch, and volume, that change from situation to situation. We incorporated explicit instructions for such nonverbal features into speech synthesis to make synthesized dialogue as much as realistic.

We synthesized a spontaneous dialogue based on a transcription of real interview. We use two systems. One of them is "Syaberinbo Portable (Syaberinbo)" from NTT Data. And another is a synthesizer (H's-Synthesizer) which is made by Hirose Laboratory in Tokyo University, the same group of this project.

Both of the synthesizers can not represent overlap of utterances. Therefore we removed transcriptions which overlapped later before synthesis. There was no way to synthesize for some nonverbal features, because each synthesizer has some limits. Syaberinbo can not set pitch by command. But we can switch female and male voice and can be changed tempo and loudness in 8 level and 16 level for each. Syaberinbo need accents and pauses information. H's-Synthesizer can not represent speakers and can not change tempo, loudness and pitch.

Both synthesizers can not represent voice qualities and make prosodic control based on the rules programed.

Original dialogue is an interview between female interviewer and male interviewee. Duration of it is about 40 seconds. Both synthesizers used in this experiment are not designed to synthesis spontaneous dialogues. Syaberinbo is designed for data reading. But, when prosodic boundaries (bunsetsu) and accent information are given, synthesized sound become almost spontaneous. As we add more nonverbal information to reading data, we feel synthesized sound become more spontaneous.

5. A Method for Temporal Measurement of Spontaneous Speech : TEMAX[3]

We reports an algorithm developed for measuring speech speed on the basis of periodicity of syllable or stress production, i.e. the tempo or the rhythm. A sense of the tempo or the rhythm is perceptible for semi-periodic stimuli with intervals between 0.1 and 3 seconds, which corresponds approximately to a syllable rate in syllable-timed language or a stress rate in stress-timed language. The speech speed, or the rate of speech, is defined as a number of mora in a second for Japanese, which was phonological rather psychological and insufficient to be measured acoustically, provide here a solution to the difficult problem of physical measurement of sensation.

TEMAX (Temporal Evaluation and Measurement Algorithm by ks) is a signal processing procedure to evaluate the rate of speech by amplitude envelope of speech waveform. TEMAX gives a sonagraphic graph computed for the speech envelope sampled at 40 Hz, where speech wave is half-wave rectified and lowpass filtered at 20 Hz. In this graph, higher energy components make frequency bands (around 8 Hz) which corresponds to the rate of speech, because syllables are produced semi-periodically in Japanese, the isosyllabic characteristics. Speech wave is amplitude modulated about 8 mora/second. Dips of the envelope corresponds to consonants or phonemic boundaries, therefore dips within a unit of time is correlated with the rate of speech. By DFT of the envelope with 1 second window, a dominant frequency component corresponds to the rate of speech.

A mora is a prosodical term that is a conjunction of a consonant and a short vowel. One measurement of the rate of speech is achievable through the phoneme recognition, that is, phonemes are marked along the time line and the resulting phoneme (hence the mora) lengths are averaged along some intervals resulting a number of moras per second. This definition, however, is difficult to realize with automatic speech recognition technology, but possible only by hand labeling. We can recognize the rate of speech without recognizing the content of speech. We can perceive, for example, the utterance speed from the narrow band filtered speech sound. This fact suggests that our sense of a tempo can be perceivable from the envelope of the waveform.

The speaking rate is observed as a dominant spectral peak in a frequency domain, where the speaking rate is visually represented in frequency-time plane like the formantic pattern in spectrograms. The frequency and the time are scaled downward to one two hundredth of the 8 kHz sampling rate of the normal spectrogram. We could observe gray gauged monochrome patterns in the 20 Hz frequency region with one second of time window.

First, we examined with the synthesized sounds such as a stationary half wave rectified signal with a short silent gap corresponding to a consonant interval, and then gradually decreasing intervals between envelope peaks. These test signals are processed according to the described procedure to obtain spectrograms of the envelope. We could observe dark bars corresponding to the speech rates.

We examined real dialogue speech taken from TV programs. The real speech rate is measured manually by segmentation of individual phonemes. Then we computed DFTs of the envelop waves to find the frequency of the peak energy as an estimate of the speech rate. The manual estimation and the DFT estimation correlated with coefficient 0.57.

TEMAX analysis on real spontaneous speech showed often multiples of dominant frequency components which is lower than manually counted frequency of mora occurrences. This is due to the bimoraic phenomena of Japanese rhythm that is recently found in theoretical study of spoken Japanese and also in psychological experiments. Analysis of speech envelope showed that two adjacent moras connect together composing a single large peak as frequently as single-mora peaks. Therefore the speech rate in the ordinal sense is observed over the bimoraic speech rate in a TEMAX. In Japanese, the rate of speech observed in TEMAX was almost steady in read speech and monologue. The rate of speech, which shows paralinguistic characteristics, changes wider to express and to exchange feelings each other beyond words.

There are two different temporal structure of language; isosyllabism such as mora-timed Japanese and isochronism such as stress-timed English. Although the TEMAX was developed based on the mora-timed Japanese, it turned out to be useful also for other type of temporal systems such as English, because the TEMAX extracts a kind of universal characteristics. TEMAX showed that stress is periodic around 1 Hz and this isochronic characteristics is observable with the analysis window of 2 seconds' width. Furthermore, the periodicity of syllables between stress is also observable with the TEMAX showing values around 5 to 10 Hz, which varies from interval to interval wider than TEMAX of syllable-timed speech. Thus, English and Japanese, two extreme language in rhythmic structure, were successfully analyzed.

6. Conclusion

In this study, we developed methods to analyze dialogue speech. First we established an encoding framework of dialogue speech, and on which we made various measurements of consistency of descriptions between different transcribers on some conditions. As the result, we got a consistency rate about 91 % for verbal sounds and 64 % for tags. Transcribers' skill and GUI had influenced on results. Then we intuitionally evaluated transcriptions including paralinguistic features by resynthesis of dialogue speech reproduced from auto-translated control command to the speech synthesizers. We got more realistic dialogue as we set not only prosodic boundaries (Bunsetsu), but accent information, tempo and volume as well. Secondly we implemented a signal processing algorithm with sonagraphic display named TEMAX to extract tempo, one of paralinguistic features of dialogue speech. TEMAX could effectively extract tempo measured in terms of not only more/second for mora-timed Japanese, but also stress/second for stress-timed English and syllable/second for syllables between stress in English.

References

[1] C.M.Sperberg_McQueen,Lou Burnard, "Base Tag Set for Transcription of Spoken Texts", TEI P2 Chapter 34, Text Encoding Initiative, Chicago, Oxford,(1992).

[2] Kobayashi, S., Kitazawa, S.: "Consistency of Inter-Transcriber's Transcription", EUROSPEECH, 1263-1266, 1995.

[3] Kitazawa, S., Kobayashi, S., Matsunaga, T., Ichikawa, H.: "Tempo Estimation by Wave Envelope for Recognition of Paralinguistic Features in Spontaneous Speech", in Proc. ICSLP, vol. 3, 1691-1694, 1994.

[4] Related documents and materials are available through the following URL: "http://mickey.cs.inf.shizuoka.ac.jp/research/".

Keywords: dialogue speech, speech rate, spontaneous speech synthesis, TEI, TEMAX