K-100

A Study on Context Effects in Speech for the Advanced Knowledge Videoconference

ﾃｨﾃｨﾎﾟｰﾝ ﾙｰﾄﾗｯﾄﾃﾞｰﾁｬｰｸﾝ† 青木 輝勝† 安田 浩†
Thitiporn Lertrusdachakul Terumasa Aoki Hiroshi Yasuda

1. Abstract

Videoconference today mainly represents the verbal forms of information exchange. However, it consists of a lot of nonverbal expression, which can convey more complicate information. The integration of the context information of the audio and video signal has the benefit to enhance the information supporting system. This paper describes a study on the context effects in speech for the advanced knowledge videoconference.

2. Introduction

In a digitally connected world, large and diverse sets of people can now, easily and cheaply, communicate to each other in more complex activities. As a result, videoconference is more relevant than ever, and the benefits are more significant. The successful integration of audio, video and data presentation tools has increasingly contributed to the community in more comprehensively social aspect. To correspond with this expansion, the advance knowledge videoconference is proposed for the advanced communication in the future. The idea is to anticipate the users` interest based on the context information of the videoconference. Then it will automatically support the relevant information to guide and introduce the user to the new idea and interesting topic of conversation. The challenging work is to find the time of retrieval and the efficient keywords. Therefore, the fundamental knowledge of speech is required. Section 3 describes the overview and the basic idea emphasized on the prosody of pitch information. In addition to the speech characteristic, the video from the conference, which deals with the images and motion, also carries the meaningful information. It can be modified and sometimes even reverses the useful information of lexical channel.

3. Fundamental Knowledge of Speech

The speech-sounds such as vowels and consonants function are mainly to provide an indication of the identity of words and the regional variety of the speaker. Another interesting aspect of speech development is the prosody. The prosody refers to all aspects of sound systems above the level of segmental sounds. It is a parallel channel for communication, carrying some information that cannot be simply deduced from the lexical channel. Prosody [1] is used to convey

- lexical meaning (accents and tones)
- non-lexical information (intonation type: question vs. declarative sentences)
- discourse functions (focus, prominence, discourse segments, etc.
- emotion

Prosody, as expressed in pitch, gives clues to many channels of linguistic and para-linguistic information. Linguistic functions such as stress and tone tend to be expressed as local excursions of pitch movement. Intonation types and para-linguistic functions may affect the global pitch setting, in addition to characteristic local pitch excursion near the edge of the sentence (i.e. boundary tones). Intonation refers to the pitch pattern of an utterance. A falling pitch might be used for declaratives and a rise pitch for imperatives. Pitch height and loudness may be genetically encoded as signal for excitement [2].

4. Context Effects in Speech

The study on the prosodic features that affects the time of retrieval and the efficient keyword for the advanced knowledge videoconference is investigated. If the prosody carries meaning, emotion, and information,

- How is all this information carried on a shared channel?
- How is it encoded?

We propose the five parameters from the speech analysis that affect the time of retrieval and the efficient keyword as follows.

1. term frequency
2. part of speech (noun)
3. question word
4. pitch or frequency
5. intensity or loudness

In this paper, we do the analysis and focus on the last two parameters; pitch and intensity. The varying quantities help to determine the interpreting of utterance. The speech analysis of pitch and intensity of the conversation is done to describe the primary characteristic of speech that contains the keyword.

5. Experimental Results

The speech analysis of two conversations was examined to determine the characteristic of pitch and intensity of the keyword representing the content of the conversation. There are two participants in each

† 東京大学

conversation. Figures 1 and 2 show the signal waveform, pitch, and intensity (bold line) over the time of one part of the conversation 1. The graphs are presented in varying quantities in every spoken utterance. The pitch analysis is based on the following parameters.

Pitch method: Autocorrelation
Pitch max. number of candidates: 15
Pitch silence threshold: 0.03 of global peak
Pitch voicing threshold: 0.45 (periodic power / total power)
Pitch octave cost: 0.01 per octave
Pitch octave jump cost: 0.35 per octave
Pitch voiced/unvoiced cost: 0.14 Hertz

The shadow areas in Figures 1 and 2 represent the keywords of the conversation from participant 1 and 2, respectively.

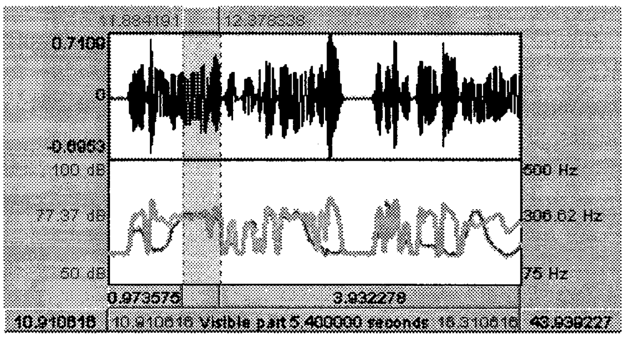

Figure 1. The speech analysis of pitch and intensity of the keyword from participant 1 of the first conversation.

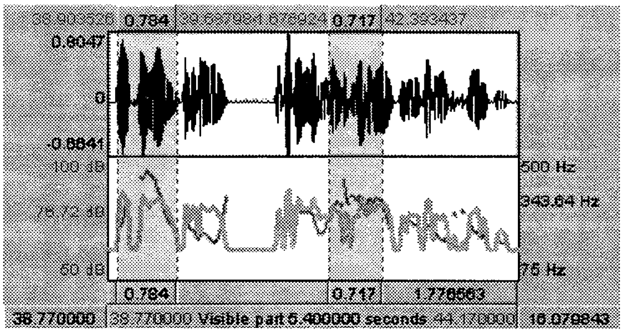

Figure 2. The speech analysis of pitch and intensity of the keywords from participant 2 of the first conversation.

The analysis data is summarized in Table 1. The average pitch and intensity of each keyword is determined and compared with the average pitch and intensity of the conversation for each participant.

where C is the n^th^ conversation
P is the n^th^ participant in the conversation
K is the n^th^ keyword of the participant
P_k is the average pitch of the keyword in unit of Hz
I_k is the average intensity of the keyword in unit of dB
P_{avg} is the average pitch of the participant in the conversation (Hz)
I_{avg} is the average intensity of the participant in the conversation (dB)
C_P is the percentage of keywords given the pitch higher than the average value
C_I is the percentage of keywords given the intensity higher than the average value

Table 1. The analysis data of pitch and intensity of keywords in the conversation.

C	P	K	P_k(Hz)	I_k(dB)	P_{avg}(Hz), I_{avg}(dB)	C_P, C_I
1	1	1	305.51	77.18	255.15, 76.39	100%,100%
	2	1	379.00	76.39	282.50, 75.05	100%,100%
		2	326.29	76.60		
2	1	1	458.14	82.90	299.80, 80.23	67%, 67%
		2	311.99	81.85		
		3	290.78	76.79		
	2	-	-	-	297.60, 79.11	-
Average						89%, 89%

The result shows that 89 percent of the keywords from three participants has the pitch and intensity level higher than the average value. Interestingly, those keywords have the pretty much higher in pitch difference than the intensity difference. However, they all have the higher level for both pitch and intensity.

6. Conclusion and Future Work

The interesting keywords usually have both pitch and intensity level higher than the average value. This implies that the high pitch and intensity might contain the keywords. However, we still need to analyze more conversations to confirm this hypothesis. The speech has the complicated structure and depends on the person. To increase the efficiency of the results, the pitch and intensity analysis need to combine with the other parameters i.e. term frequency, part of speech, and question word. In addition, the integration of the video information such as facial expression and facial direction needs to be focused to further research on the advanced knowledge videoconference.

7. References

[1] C. Shih and G. Kochanski, "Prosody and Prosodic Models," *Proc. the 7^th^ International Conference on Spoken Language Processing,* 2002.
[2] P. Baker and T. Mcenery, "Language Acquisition: Phonological and Lexical Development," http://www.ling.lancs.ac.uk/chimp/langac/LECTURE5/5home.htm.