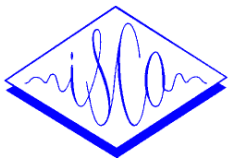

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

TEMPO ESTIMATION BY WAVE ENVELOPE FOR RECOGINITION OF PARALINGUISTIC FEATURES IN SPONTANEOUS SPEECH

Shigeyoshi Kitazawa, Satoshi Kobayashi, Takao Matsunaga and Hideya Ichikawa

Shizuoka University

Hamamatu-shi, Shizuoka, 432 Japan

ABSTRUCT

We analyze speech rate through an envelope extraction process. The process is low-pass filtering of rectified speech wave to eliminate ripples caused from pitch and vocal resonances. Speech wave is amplitude modulated about 8 mora/sec.. Dips of the envelope correspond to consonants or phonemic boundaries, therefore dips within a unit time is correlated with the rate of speech. We measured the rate of speech from an interviewing between a female interviewer and a male interviewee. Speech data analysed consists of 7 utterances of the man and 6 utterances of the lady with durations of 2 to 7 seconds. Same utterances were labeled manually for locations of individual phonemes. Manually computed rate excluding pauses is faster than averaged one. By DFT of the envelope, a frequency component of the rate of speech is avilable and have shown to be correlated with the manual rate at the coefficient of 0.57.

I. INTRODUCTION

A lot of information is communicated nonverbally in spontaneous speech. Nonverbal communication includes vision, smell, touch, facial expression, gesture and others. Nonverbal information about voice is called paralanguage[1]. We try to measure the paralinguistic features by acoustic analysis. Paralanguage consists of vocalizations and voice qualities. Vocalization consists of semilexical one(interjection, etc.) and nonlexical one(cough, laugh, etc.).

Voice qualities consist of loudness, pitch range, speed of utterance, rhythmic qualities, pause and overlapping in conversational situations. These are very tightly connected with the prosodical and semilexical features.

Spontaneous speech is describable based on the TEI labeling method, where paralinguistic features is defined[2]. Some of these features are measurable, while some are not. In this paper we consider tempo (speed of utterance), which is one of the measurable feature. We try to define the rate of speaking, and to give means to measure it.

II. TEMPO AS PARALINGUISTIC FEARURE

The rate of speech is one of important feature of the paralinguistic features. For example, slowly speaking speech generally implys important part that speaker wants to put emphasis on. Therefore, measurement of the rate of speech will show flow of conversation.

Japanese words consist mainly of CV syllables. A (C)V syllable corresponds to one "mora", with two major exceptions. One is the syllable final nasal, seen in words like "hon", the other is the first part of a geminate consonant, as in "sippai" ,"itta". Such a consonantal unit is considered equivalent to a CV syllable in terms of the rhythmic unit known as the "mora". These rhythmic unit would be equivalent in length of utterance. We will measure the rate of speech by this definition. The rate of speech is defined as the number of "mora" per a unit of time.

III. ENVELOPE EXTRACTION

The sense of tempo is perceivable from rhythmic swing of loudness. To measure the tempo in spontaneous speech, we need to get wave envelop.

Usually the tempo is 8 mora/sec. In order to filter out the high frequency components of pitch and formants, and in order to maintain the shape of the envelope, the low-pass filter needs bandwidth of about 10 times of max frequency. So we design low-pass filters whose cut off frequency is 80 Hz.

We compared sample-type FIR filters of three different cut off frequencies; 80, 100 and 120 Hz. Then the envelope extracted by filter of 80 Hz was considerably better than the other. So we adopted the 80 Hz low-pass filter. Following figure show the characteristics of the filter.

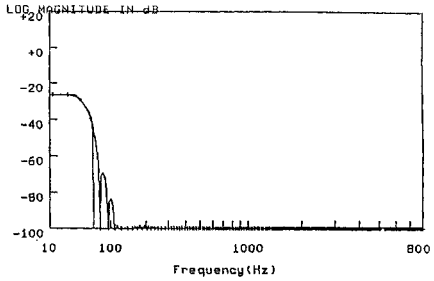

Fig.1 The characteristics of the filter.

The process to extract the envelope is shown as follows. First speech wave is converted into the absolute value of input speech wave. Second, to obtain the envelope this wave form is low pass filtered. This process is shown Fig.2.

The counts of dips approximately correspond to the number of phonemes within an utterance. The depth of dips is large around vowel-to-consonant transitions, on the other hand it is shallow around consonant-to-vowel transitions. A cluster of consonant-vowel sequence compose a mountain like envelope shape as seen is Fig.2. This cause a strong component of the period of that consonant-vowel cluster, a "mora" and that period can be detected as a frequency component by DFT.

IV. SPONTANEOUS SPEECH DATA

We try to remove element of high frequency components by low-pass filtering to extract peaks and dips of speech power, and to measure tempo by locating the start point of phonemes. The periodicity of the envelope wave within the Hamming window is extracted by DFT as a peak component, which is taken as an approximate value of the speaking rate, mora per second. And we supposed these peaks approximate the value of tempo. To show that there is correlation between the rate of speech and the measure of DFT, we explored spontaneous speech taken from a TV program of interviewing between a female interviewer and a male interviewee. Speech data analyzed consists of 7 utterances of

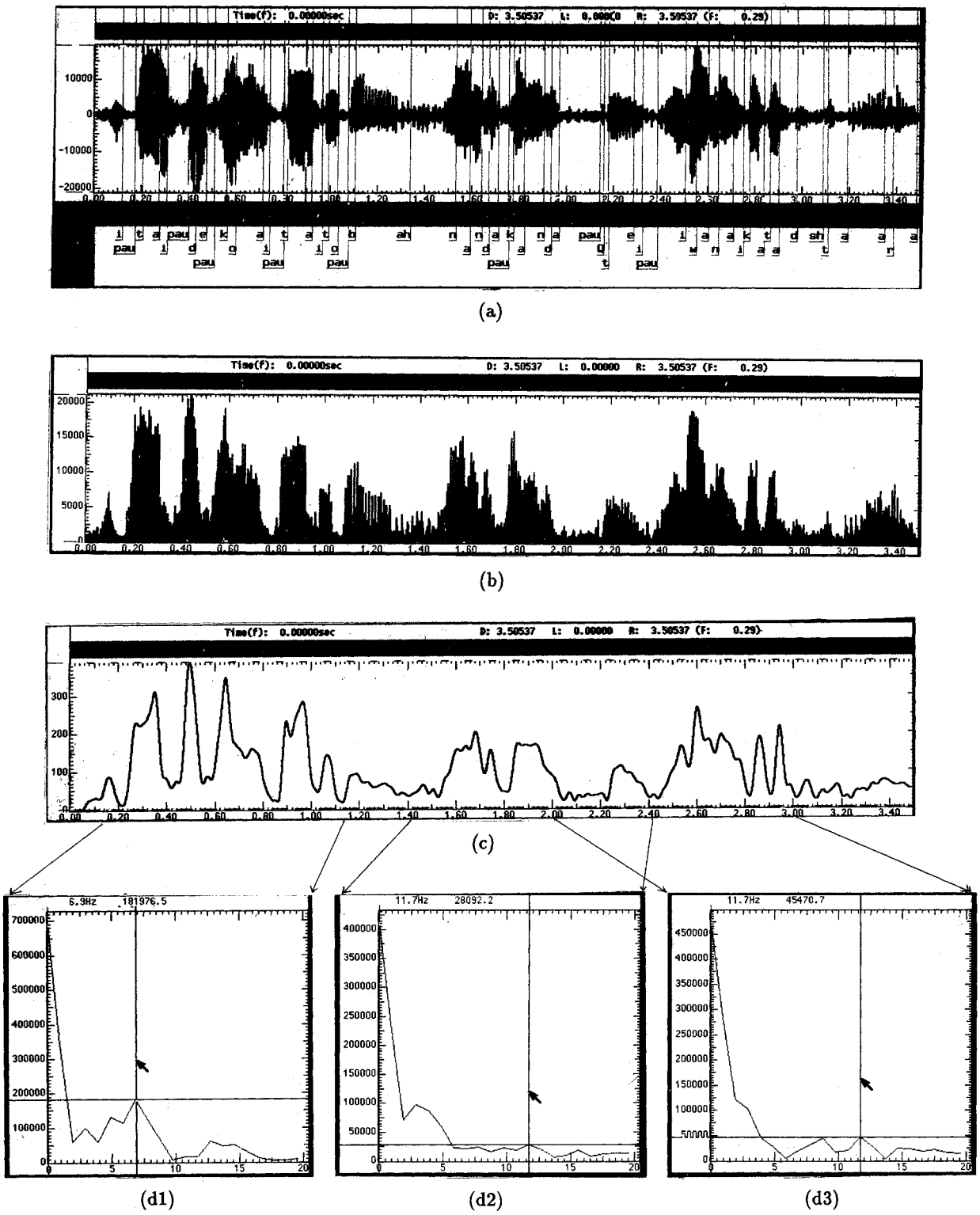

Fig.2. Signals and spectra to show internal processing of a spontaneous speech which appeared as Male 1 in Tab.1. (a) input speech wave with labels of phonemes, (b) full wave rectified speech wave, (c) low pass filtered output of the rectified signal showing dips approximately corresponding to phoneme boundaries and larger swings corresponding to CV syllable or mora, and (d1, d2, d3) DFT's of head(d1), middle(d2), and tail(d3) of the envelope to extract speech rate as a peak frequencies indicated by cross-hair cursor in the figures.
Number of moras and rate of mora can be counted from the labeled indexes.

the man and 6 utterances of the lady with durations of 3 to 7 seconds. Sampling frequency is 8000 Hz. Transcriptions of speech data is shown in Tab.1.

To compare measurements with real tempo in speech data analyzing, we must check it previously by labeling. Labeling is executed with listening and looking at speech wave form as seen in Fig.2(a).

V. RESULTS OF MEASUREMENTS

Tab.2 shows details of input speech data and the number of dips of envelope. The rate of speech estimated globally as a number of mora devided by the duration of an utterance.

Since the rate of speech fluctuate within an utterance, and since an utterance includes pauses, voiced or unvoiced, the *global rate of speech* is not suitable to analyse in detail the spontaneous speech. Therefore we need more local measurements of the rate of speech. For this purpose, we prepared two measurements: one is the labeled rate of speech and the other is the computed rate of speech.

The *labeled rate of speech* is calculated through manual labeling of individual phoneme. Durations of each phoneme and then durations of each "mora" is obtained with consideration of Japanese phonology. The inverse of this mora duration is the most detailed rate of speech. This value is averaged along several moras to smooth fluctuations away.

The *computed rate of speech* is calculated as a DFT component of the envelope of the utterance as we explained earlier part of this section.

Usually local measurements of the rate of speech are faster than global rate of speech, because the former excludes pauses, semilexical and non-lexical vocalization.

The column of DFT in Tab.2 is computed with Hamming window size of 1.0 sec.. The rate of speech measured at the head, middle and tail of the utterances. *Head* point is about 0.5 sec from the top, *middle* is the center, and *tail* point is about 0.5 sec before the end.

Fig.3 shows correlation of the manually labeled rate of speech to the global rate of speech, and Fig.4 shows correlation of the computed rate of speech (DFT) to the global rate of speech. Points scatter upper left of the graph showing faster rate of measurements. Fig.5 shows correlation between the computed rate of speech and the labeled rate of speech. The correlation coefficient is 0.57 showing good correlation.

VI. DISCUSSION

As results shown in previous section, measurements of the tempo by extracting dips of envelope and by DFT are obtained.

These measurements are correlated well with the global rate of speech as well as the labeled rate of speech. And the measurements matche to our intuition. We think that in these ways we can measure tempos automatically.

To measure tempo more exactly, those parts with small fluctuations of the envelope, such as long silent pauses, inhaling noises, which appear frequently in spontaneous speech, should be processed appropriately. Amount of speech data processed is small for the moment and we must collect many data to have more reliable measurement.

VII. CONCLUSION

We analyzed speech rate through an envelope extraction process. We used low-pass filtering to detect temporal pattern and DFT to estimate speech rate. A simple signal processing could obtain reliable approximate of speech rate from wide ranges of spontaneous speeches. And these measures can be used to recognize paralinguistic features as well as to segment individual phoneme during speech recognition.

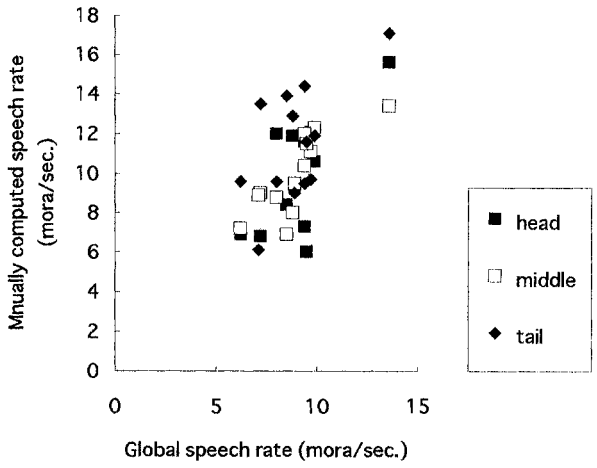

Fig. 3. Manually computed speech rate correlated with global speech rate.

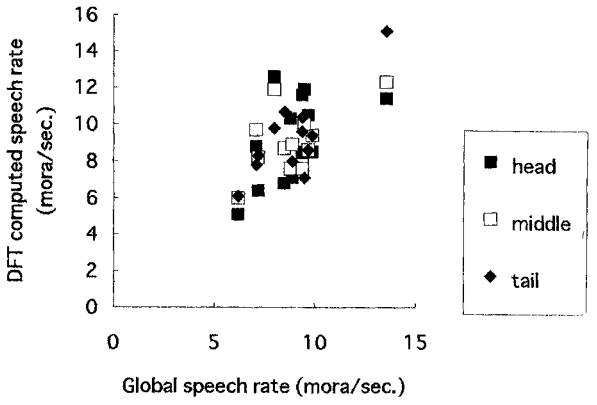

Fig. 4. DFT computed speech rate correlated with global speech rate.

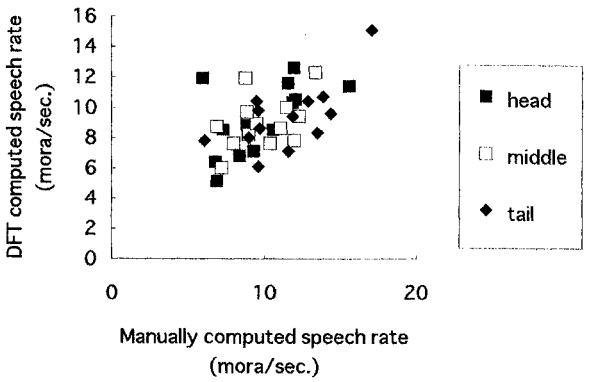

Fig. 5. DFT computed speech rate is correlated with manually computed speech rate.

REFERENCE

[1] George L. Trager:
"Paralanguage: A First Approximation"
Studies in Linguistics,13,1-11,1958.

[2] C.M.Sperberg-McQueen,Lou Burnard:
"Transcription of Speech" TEI P3 Chapter 11,
Text Encoding Initiative,Chicago,Oxford,1993.

Tab.1 Transcriptions of samples of utterances excised from a TV interview.

Male1	itai dokogaitaitoka nandakandaiwanai katadesitakara
Male2	soudesuneh anoh nagaisihnnankattedatoh zuttomottenakyanannaindeh
Male3	owarinohouni naruto tega sibiretekityaundesuyoh
Male4	eh demoitijikan itijikanguraidesukaramah nijikanwa sitijika osokutomo hatijiguraidesune
Male5	ndeh kujiguraikara norihajimeteh nijiguraimade desukaneh
Male6	notte desaigo mata tyokotto umano teiresite kaerun
Male7	nitiyoubiga yasumikato omottandesukedoneh umanokeikoga arimasite tyoubaga
Female1	ma imawa gendaiteki gendaigekio yaritaimonodato omotterassyarusoude gozaimasuga
Female2	demo sonosanninno ohdisyonno hokanohitotatimo minna neoagezuni soredemo
Female3	utini kityattandesuyoneh hagakimitainamonde sorede nangatu nannitini oatumarikudasaitteittara moukonaihitoga iruwakejanai
Female4	uhn dakara souyunotte sugoku yokuwakarujan annanikataku oyakusokusitanonitteh
Female5	ma yoru gohanga owatte otyato nitibutte yukotonandesukedo sokode mata anosehza
Female6	taihhendesu sorede oyasumiga itinitimonaitte yukotowa nitiyoubiha

Tab. 2. Analyzed data of the utterances in Tab. 1.

	A	B	C	D	E=B-D	F	G=C/A	manual computed			DFT computed (Hz)		
sample	sec.	phnms	moras	dips		%	rate	head	middle	tail	head	middle	tail
M1	3.505	48	28	50	-2	4.2	8	12.0	8.8	9.6	12.6	11.9	9.8
M2	4.000	48	34	54	-6	12.5	8.5	8.4	6.9	13.9	6.8	8.7	10.7
M3	1.685	28	23	24	4	14.3	13.6	15.6	13.4	17.1	11.4	12.3	15.1
M4	4.977	69	44	61	8	11.6	8.8	11.9	8.0	12.9	10.3	7.6	10.4
M5	4.161	47	30	57	-10	21.3	7.2	6.8	9.0	13.5	6.4	8.2	8.3
M6	3.537	39	25	57	-18	46.2	7.1	6.9	7.2	9.6	5.1	6	6.1
M7	4.257	55	38	48	7	12.7	8.9	9.3	9.5	9.0	7.1	8.9	8
avg.M	3.731	47.7	31.7	50.1	9.20*	17.5	8.9	10.1	9.0	12.2	8.5	9.1	9.8
F1	4.257	62	42	58	4	6.5	9.9	10.6	12.3	11.9	8.5	9.4	9.4
F2	3.697	66	36	53	13	19.7	9.7	12.1	11.1	9.7	10.5	8.6	8.6
F3	6.929	102	65	94	8	7.8	9.4	11.6	12.0	14.4	11.6	7.8	9.6
F4	4.273	62	41	56	6	9.7	9.6	7.3	10.4	9.5	8.5	7.6	10.4
F5	5.521	67	39	81	-14	20.9	7.1	8.9	8.9	6.1	8.8	9.7	7.8
F6	3.665	54	35	49	5	9.3	9.5	6.0	11.5	11.6	11.9	10	7.1
avg.F	4.724	68.8	43	65	9.18*	12.3	9.2	9.4	11.0	10.5	10	8.9	8.8

A : duration, B : phonemes, C : moras, D : dips detected, E=B-D : difference, F=|E|/Bx100 : percent error, G=C/A : global speech rate in mora/sec., manual computed : local rate of speech excluding pauses, DFT computed : computationally estimated speech rate. head : speech rate in beginning1 sec. middle : speech rate in the middle 1 sec. tail : speech rate in 1 sec. at the end.
*: standard deviation of differences.