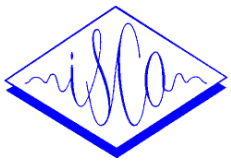
ISCA Archive
http://www.isca-speech.org/archive



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

A NEW APPROACH TO COMPENSATE DEGENERATION OF SPEECH INTELLIGIBILITY FOR ELDERLY LISTENERS

Akira Nakamura, Nobumasa Seiyama, Atsushi Imai, Tohru Takagi and Eiichi Miyasaka
NHK Science & Technical Research Laboratories,
1-10-11, Kinuta, Setagaya-ku, Tokyo 157, Japan

ABSTRACT

This paper presents a new hearing aid system intended to compensate the degradation of perceptual functions in the central auditory pathways which can be found typically with elderly people, while conventional hearing aid systems are effective only for conductive hearing impairments. A typical type of such functional deterioration is the decrease of rate of processing speed for identification of speech and effective cognitive capacity. The new system is designed to compensate it by slowing the input speaking rate instead of a direct processing of the related auditory functions. This system enables a user to convert a speaking rate as desired by him/her-self on real time, with invariance in pitch as well as small impairments in quality.

INTRODUCTION

According to the population statistics announced by the Institute of Population Problems of the Ministry of Health and Welfare in Japan, people aged 65 or more comprised 12.9% of the entire Japanese population in 1992. The Institute predicts that the population of the elderly will continue to increase rapidly, comprising more than 20% of the entire population in 2010 and comprising 24% by 2015, to double the present figure. In about 20 years, one out of four radio listeners or TV viewers will be a person aged 65 or more. It will be urgently necessary to provide broadcasting service that accommodates older listener's hearing ability.

In the past several decades, speaking rate in broadcasts has accelerated in Japan. Average speaking rate in broadcasting was 340 mora/minute several years ago. However, recent statistics indicate that the rate has often reached 450 to 570 mora/minute. Some TV personalities speak at the rate of 770 mora/minute.

It is well known that older listeners have some degree of hearing impairments, such as increased absolute thresholds of audibility and decreased intelligibilities of speech uttered rapidly or in noise. It has been shown that some degree of functional deteriorations can observed in every layer of auditory systems for older individuals (sensory neural pathways, central auditory pathways, etc.)[1,2]. These deteriorations are hypothesized to decrease the rate of processing speed of identification of speech and to decrease effective cognitive storage capacities, and therefore make identification of speech uttered rapidly more difficult for elderly individuals. It is, however, difficult to compensate physiologically these deteriorated functions. We intended to compensate the decreased intelligibility of speech uttered rapidly by processing the speaking rate of input speech with small impalements.

The new developed hearing aid system can enable a user to control of speaking rate to which he/she prefers, with invariance in pitch, with small impairments, and on real time.

1. ALGORITHM[3]

The standard liner predictive coding (LPC) is often used as one of the methods for speaking rate conversion, maintaining the pitch of an original voice. Using the method, however, it is difficult to synthesize output speech naturally like an original voice. To produce converted speech maintaining naturalness, TDHS[4] (Time Domain Harmonic Scaling) method has been proposed. In this method, however, when the extension rate of a speech is close to 1.0-times, the length of the processing window becomes longer comparing to the pitch period. Therefore, if the pitch period drifts, the voice quality tends to deteriorate. So, the following new algorithm has been adopted for the speaking rate conversion.

1.1 ANALYSIS

Fig.1 illustrates the block diagram of the speaking rate converting algorithm on this system. An input speech waveform is classified into voiced-, unvoiced- and silent-portions. This classification is made by frame-synchronous analysis including some conventional techniques using power contour, zero-crossing rate, and auto-correlation analysis. For voiced portions, pitch extraction and segmentation are performed as follows; By auto-correlation analysis with multi window lengths, candidates of pitch periods around a comparably stable position within a voiced portion are extracted in every window length, and most likely one can be statistically selected as primary pitch period from them. Referring the primary pitch period, pitch segmentation over a voiced portion is carried out. It is done based on peak picking of the sinusoidal-like waveform obtained by low-pass filtering. The cut-off frequency is decided frame by frame as a little higher than the average pitch frequency which is estimated around the extracted primary pitch frequency. The above analytic procedures are performed automatically.

1.2 EXTENSION METHOD FOR VOICED SPEECH[3]

Fig.2 illustrates the method of extension of a voiced por-

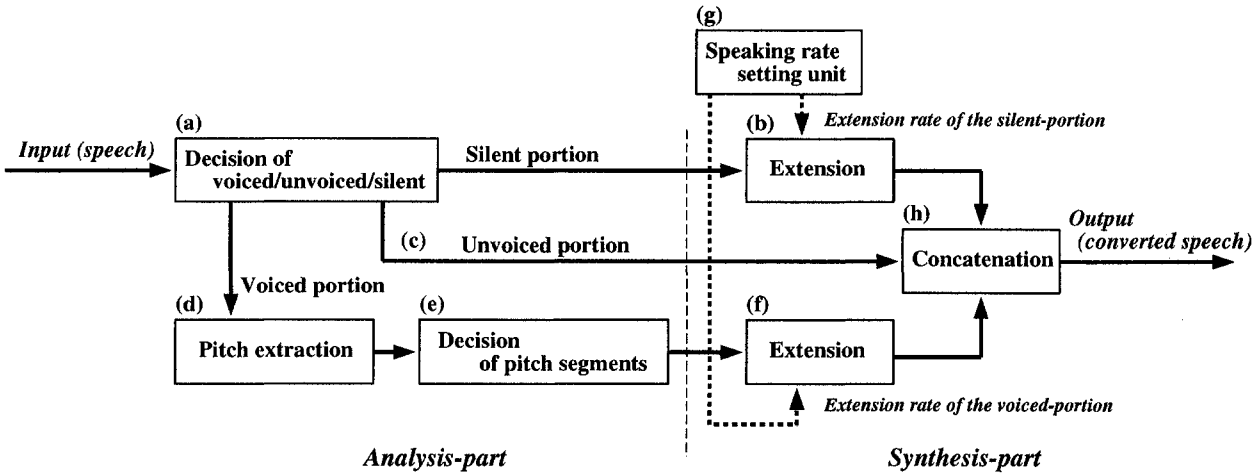


Fig.1 A block diagram of the speaking rate converting algorithm

tion S (**a**, **b**, **c**, **d**, **e**, **f**, **g**, ...), where **a**, **b**, **c**, ... are pitch segments, and **Pa**, **Pb**, **Pc**, ... are pitch periods. Simple temporal uniform extension of S produces S'(***a'***, ***b'***, ***c'***, ***d'***, ***e'***, ***f'***, ***g'***, ...) with lowered voice pitch. S" is the extended version of the voiced portion by this method, maintaining the voice pitch of the original S.

(1) The first pitch segment **a** is always selected as the first segment of S".

(2) For the second pitch segment of S", either **a** or **b** should be selected.

(2-1) The overlapping rate (**x**) of **a** and **a'** is calculated. In this case, **x=Δa/a** (**Δa=*a'*-a**).

(2-2) The overlapping rate (**y**) of **b** and **b'** is calculated. In this case, **y=Δb/b** (**Δb=(a+b)-*a'***).

(2-3) if **x>=y**, **a** is selected, else , **b** is selected. In this case, **y** is larger than **x**, so that the pitch segment **b** is selected.

(3) For the third pitch segment, either **b** or **c** should be selected. In this case, **c** does not overlap with ***c'*** temporally, so that the pitch segment **b** is selected.

(4) For the fourth pitch segment, either **c** or **d** should be selected. In this case, **c** overlaps with ***c'*** completely, so that the pitch segment **c** is selected.

The above process is continued until the total length of the extended voiced portion exceeds the length of the uniform extension S'. As a result of it, the extended version S" of the voiced portion can be obtained with maintaining the voice pitch of the original speech.

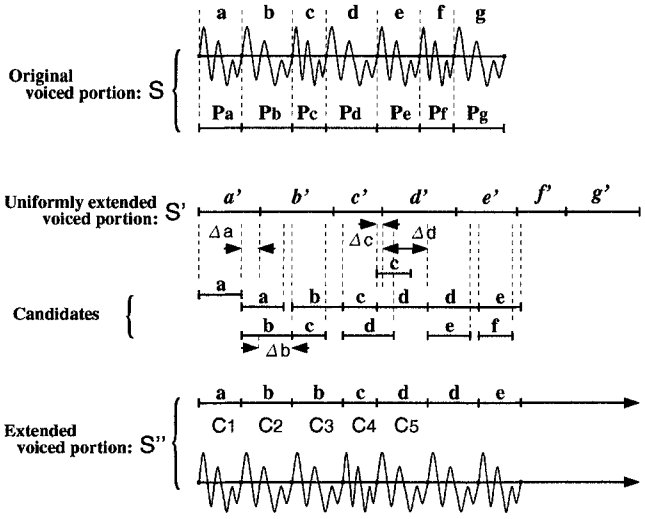


Fig.2 An example of extension of a voiced portion

2. HARDWARE

Fig.3 illustrates the hardware block diagram of the real time speaking rate converting system. Each box consists of a Transputer module (parallel processing LSIs: TRPM), and DSP (digital signal processors) for executing the assigned signal processing algorithm at a high speed enough to maintain continuity of the output speech. An appropriate sequential-parallel connection of these TRPMs and DSPs enables to convert a speaking rate on real time. The analysis-part shown in Fig.1 is executed through three TRPM · DSP units (TRPM-1～3 · DSP-1～3), while the synthesis-part is executed only through the TRPM-4 · DSP-4. A function of networking and buffering is mainly allo-

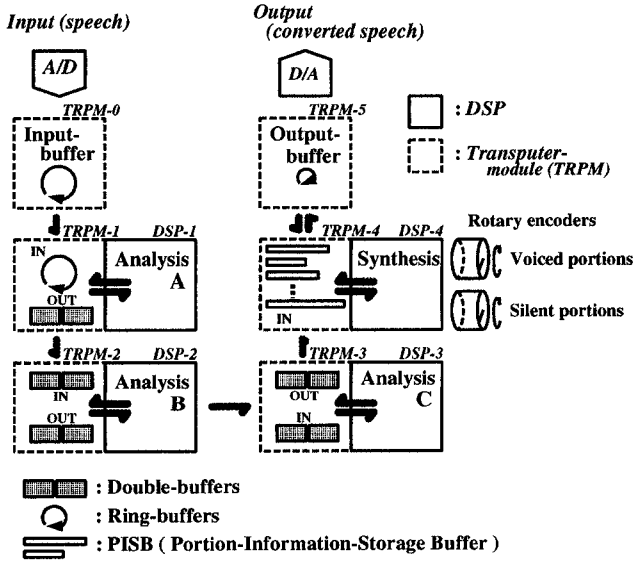


Fig.3 Hardware of a real time speaking rate converting system

cated to the TRPMs, while calculation of signal processing is done by the DSPs.

2.1 INPUT

TRPM-0 carries out buffering on the speech data digitized by an A/D converter with 16-bit resolution at a rate of 16 kHz sampling.

2.2 ANALYSIS-PART

Each TRPM has a double-buffer for the purpose of increasing the processing speed. In the Analysis units A～C, decimation of input speech data is carried out to reduce the amount of processing required in pitch extraction and segmentation for each voiced portion. The output through TRPM-3 includes the starting and the ending points of each pitch segment, the number of pitch segments in each voiced portion as well as the original digitized waveform data.

2.3 SYNTHESIS-PART

This system is equipped with two rotary encoders with 8-bit resolution in order that silent portions and voiced portions can be independently controlled by a user. Each encoder converts a rotated angle into the corresponding rate of silent- or voiced-portions. The range of the rate of extension is set from 1.0 (original) to 1.6 for voiced portions, and is set from 1.0 (original) to 3.0 for silent portions respectively based on the results of hearing tests[15].

In the DSP-4, the voiced- and silent-portions are extended according to the input from the rotary encoders, while the unvoiced portions are not at all processed. Then an appropriate concatenation of the extended voiced-, silent- and the unvoiced-portions is carried out, resulting in an extended speech with small impairments.

2.4 PISB

TRPM-4 includes a function of a Portion-Information-Storage Buffer (PISB) which enables the rate of speech to change quickly in response to a listener's operation. It stores (1) the starting and ending point of each silent-, unvoiced- and voiced-portion, (2) the starting and ending point of each pitch segment, and the number of the pitch segments in each voiced portion and (3) the original speech data obtained from DSP-3. TRPM-5 is the output buffer and its capacity is set to be only 320-bytes, corresponding to 10ms-speech data with 16-bit resolution at a rate of 16 kHz sampling. TRPM-5 monitors the amount of speaking rate converted data stored in the output buffer. If the amount decreases fifty percent, DSP-4 gets a request to read new data from the PISB (TRPM-4) successively, and then DSP-4 concatenates them. This process enables the speaking rate to change quickly whenever a user wants to convert.

3. DISCUSSION AND CONCLUSIONS

3.1 QUALITY ASSESSMENT TESTS FOR THE CONVERTED SPEECH

The quality of a speaking rate converted voice by using a conventional method is usually deteriorated and the articulation decreases with the rate of extension. Even listeners with normal hearing become hard to comprehend the converted speech if the converted voice is distorted. We conducted the articulation test using the following Japanese syllables. The original 100 Japanese syllables spoken by a male and a female skilled announcers consist of 5 individual vowels and the pair of 25 consonants followed by these vowels. Each syllable was converted at a speak-

(1) Extension rate of voiced portion : 1.0 (Original)

T \ R	p	t	k	h	ky	hy	b	d	g	r	by	ry
p	99.93	0.07										
t		100.0										
k			100.0									
h				100.0								
ky					99.98	0.02						
hy						100.0						
b							99.98	0.02				
d								100.0				
g									100.0			
r										100.0		
by											100.0	
ry												100.0

(2) Extension rate of voiced portion : 1.2

T \ R	p	t	k	h	ky	hy	b	d	g	r	by	ry
p	99.95	0.05										
t		100.0										
k	0.05	0.02	99.93									
h	0.02			99.98								
ky					100.0							
hy						100.0						
b							100.0					
d							0.02	99.98				
g									100.0			
r										100.0		
by											100.0	
ry												100.0

(3) Extension rate of voiced portion : 1.4

T \ R	p	t	k	h	ky	hy	b	d	g	r	by	ry
p	99.95	0.05										
t		100.0										
k			100.0									
h				100.0								
ky					100.0							
hy						100.0						
b							100.0					
d								100.0				
g									100.0			
r										100.0		
by											100.0	
ry												100.0

***T*: Transmitted- , *R*: Received- monosyllables**

Table.1 Confusion-Matrix for the consonants when the voiced portion was extended by 1.0 ～ 1.4 times (for the male announcer)

ing rate of 1.0, 1.2, 1.4-times by using our system. The subjects were 3 females in their 20s with normal hearing. The converted syllables were randomly and presented 20 times to every subject in a sound proof room through a loudspeaker with approximately 70dB(A).

Table 1 (for the male announcer) is the confusion-matrix for consonants when the voiced portion was extended by 1.0~1.4 times. The results for the vowels were omitted in the table because the scores were 100%. Table 1 shows that the scores of the articulation were almost 100% except 12 consonants showing lower score than 100%. The details of these confusions are as follows;

(1) Although the confusion score was 0.05% in the case of /p/→/t/, a similar confusion was also observed in the original voice. It can be considered that the confusion is not produced by this system.

(2) When the voiced portion is extended by 1.2 times (Table 1-2), the confusion becomes 0.02 to 0.05% in the case of /k/→/p/, /k/→/t/. When the speaking rate is increased to 1.4 times, no confusion was observed.

The results show that the speaking rate converted voice can hold enough quality to the original voice.

3.2 INTELLIGIBILITY FOR THE ELDERLY LISTENERS

Using this system, an experiment for intelligibility of the slowed speech was conducted for the elderly. The original material is an approximately ten-second Japanese news sentence spoken by a skilled male announcer at a rate of roughly 9.4 mora/sec. The stimuli were made by converting the rate of 1.0, 1.2, and 1.4 times. The evaluation was done on three categories: "easy to hear," "ordinary," and "difficult to hear."

The subjects were 15 persons, consisting of seven elderly listeners with presbycusis and eight young listeners with normal hearing. The stimuli were provided through a loudspeaker with MCL (Most Comfortable Level).

Fig.4 shows the average data. The left part of the figure shows the results for young listeners, while the right part shows the elderly listeners' results. The figure indicates the following trends.

(1) In the case of 1.2-times extension of the voiced portions, the elderly group gave the same evaluation as in the case of the original rate, while the evaluation for the young group increased to 62.5% from 37.5% for the original speech.

(2) In the case of 1.4-times extension, the score for the elderly group nearly doubled to 85.7% from the 42.9% for the original speech, while the young group commented that "it becomes unnatural" because of the excessively slow speaking rate.

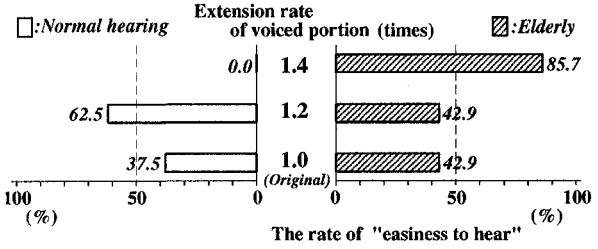


Fig.4 The rate of "easiness to hear" for speaking rate converting voice

These results indicate that even the young listeners feel easy to hear when the speaking rate is extended to 1.2 times, while the elderly listeners prefer further more extended rate.

4. SUMMARY

This paper presents a real time speaking rate converting system for the purpose of compensating degradation of intelligibility of speech uttered rapidly. This system can be applied not only to Japanese spoken language but also other spoken foreign languages, such as English, German, French, Chinese and so on.

When this system is used for speech accompanied visual image, temporal discrepancy between the converted voices and visual images occur. At present, we have already developed an advanced real time speaking rate conversion system installing a new algorithm. The new algorithm includes absorption of temporal extension produced in converted speech by temporally varying the speaking rate along a curve calculated referring to the change of the fundamental frequency in a sentence.[6,7] We have also discussed on inarticulate impression of converted speech at a very slow rate and are proposing an advanced new algorithm to reduce the inarticulateness.[8] Detail are now being compiled in a different paper.

ACKNOWLEDGMENTS

The authors wish to thank Prof. S.Funasaka (Department of Otorhinolaryngology, Tokyo Medical College), Prof. H.Ono (The National Center for University Entrance Examination), and Dr. H.Seki (Department of Computer Science, Chiba Institute of Technology) for their clinical tests by using this system.

REFERENCES

[1] E.W.Herold, "Audio for the Elderly", J.Audio.Eng.Soc., Vol.36, No.10 (1988)

[2] T.Satoh and T.Adachi, "Identification of time required by elderly individual for the perception of vowels", Audiology Japan, vol.31, pp.737-743 (1988)

[3] A.Nakamura, N.Seiyama, R.Ikezawa, T.Takagi, and E.Miyasaka, "Real time voice speed converting system for elderly people", Procs. of IC-ASSP'94, pp.225-228 (1994)

[4] D.Malah, "Time-domain Algorithms for Harmonic bandwidth Reduction and time scaling of speech signals", IEEE trans. Acoust., Speech, Signal processing, vol.ASSP-27, pp.121-133 (1979)

[5] A.Nakamura, N.Seiyama, R.Ikezawa, T.Takagi, and E.Miyasaka, "Real time voice speed converting system without impairment in quality", Trans. IECE, SP92-55, pp.41-48 (September 1992)

[6] R.Ikezawa, A.Nakamura, N.Seiyama, T.Takagi, and E.Miyasaka, "A method of absorbing temporal enlargement of speech lengths in the voice speed converting system for elderly", Trans. IECE, SP-56, pp.49-56 (September 1992)

[7] A.Imai, A.Nakamura, N.Seiyama, T.Takagi, and E.Miyasaka, "Real time methods for absorption of temporal enlargement of speech introduced by the Speech Rate Converting System", Procs. ASJ Autumn Meeting, pp.361-362 (October 1993)

[8] N.Seiyama, T.Takagi, and E.Miyasaka, "Improvement of Naturalness on Speech Rate Converting Algorithm: Handling of Glides in Voiced Speech", Procs. ASJ Autumn Meeting, pp.299-300 (October 1993)