Speech Visualization by Integrating Features for the Hearing Impaired

Akira Watanabe, Shingo Tomishige, and Masahiro Nakatake

Abstract—**This paper describes development of a new speech visualization system that creates readable patterns by integrating different speech features into a single picture. The system extracts the phonemic and prosodic features from speech signals and converts them into a visual image using neither speech segmentation nor speech recognition. We used four time-delay neural networks (TDNN's) to generate phonemic features in the new system. Training of the TDNN's using three selected frames of eight kinds of acoustic parameters showed significant improvement in the performance. The TDNN outputs control the brightness of patterns used for consonants, that is, each of the consonant-patterns is represented by a different white texture whose brightness is weighted by the output of a corresponding TDNN. All the weighted consonant-patterns are simply added and then overlaid synchronously on colors due to the formant frequencies. When this is done, phonemic sequences and boundaries manifest themselves in the resulting visual patterns. In addition, the color of a single vowel sandwiched between consonants looks uniform. These visual phenomena are very useful for decoding the complex speech code, which is generated by the continuous movements of speech organs. We evaluated the visualized speech in a preliminary test. When three students read the patterns of 75 words uttered by four males (300 items), the learning curves showed a steep rise and the correct answer rate reached 96–99%. The learning effect was durable: after five months of absence from the system, a subject read 96.3% of the 300 tokens in a response time which averaged only 1.3 s/word.**

Index Terms—**Feature extraction, reading test, speech visualization.**

I. Introduction

INFANTS grow up hearing environmental sounds including speech. They learn talking by hearing and learn intensively reading and writing after reaching school age. Since hearing impaired children receive insufficient sound information, it is often hard for them to talk fluently. Moreover, cumulative insufficiency of sound information might cause delays in general learning.

In this paper, we propose a new speech visualization that is regarded as a kind of phonetic notation to be represented together with prosodic features. Speech visualized by the proposed method is one of the multipurpose media by which hearing-impaired children might develop their ability to understand speech, to talk, to improve hearing and even to read text.

Studies of speech visualization will contribute not only to developing the practical aids for the hearing impaired, but also to making clear the analogy and contrast between auditory and visual perception of speech. If the visualized speech, for example, is readable intuitively and freely, it will suggest that visual decoding of speech will have been achieved independently of the processing peculiar to hearing, like decoding coarticulation effects. One of the ultimate aims in this research is not to acquire alphabetical notation from speech signals, but to create a visual image which can be understood as easily as heard speech with respect to segmental and suprasegmental information.

Historically, one of the epoch making systems for speech visualization was "visible speech" [1], which is presently used for analysis of the sound spectrogram.

After that, "correlatogram" [2] using autocorrelation functions, "intervalgram" [3] using zero-crossing rates and "wave collation visual speech display" [4] using pitch-synchronous methods were proposed. Some modified representations of the sound spectrograms were also reported [5], [6].

The readability of speech spectrograms had been discussed until about 1984. In particular, systematic experimental results, in which a researcher (the expert reader named VZ) acquired an excellent ability after he had spent between 2000–2500 h reading spectrograms, had a strong impact on many speech researchers [7]. According to the report [7], the reader was asked to read spectrograms of 17 normal and six anomalous sentences consisting of 5–8 words each, and 45 words that had been uttered by two talkers. His reading ability was excellent despite reading the almost unknown materials. However, it seems that numbers of readers and talkers in the tests are too small to confirm generality of the results. In addition, as an important index to intuitive representation, there is no report of the response time that he spent for judgement.

On the other hand, eight subjects who were not expert but ordinary readers participated in the tests to read 50 monosyllabic English words. After the subjects had learned how to read 50 specific words, which had been pronounced by a few speakers, they tried to read 50 new words by another speaker. The correct answer rate for the new words, which they had never directly learned, was only 6% for words and 34% for phonemes [8]. This result may have originated from both a shortage of learning experience and the fact that the deformation of visual patterns, which is a result of complex encoding based on coarticulation, makes speech spectrograms hard to read as pointed out by Liberman *et al.* [9].

Manuscript received July 13, 1998; revised August 16, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dennis R. Morgan.

A. Watanabe is with the Department of Computer Science, Faculty of Engineering, Kumamoto University, Kumamoto, 860-8555, Japan (e-mail: watanabe@cs.kumamoto-u.ac.jp).

S. Tomishige is with the Oki Electric Industry Co. Ltd., Tokyo, Japan.

M. Nakatake is with the Fujitsu Oita Software Laboratory Ltd., Oita, Japan.

Publisher Item Identifier S 1063-6676(00)05178-6.

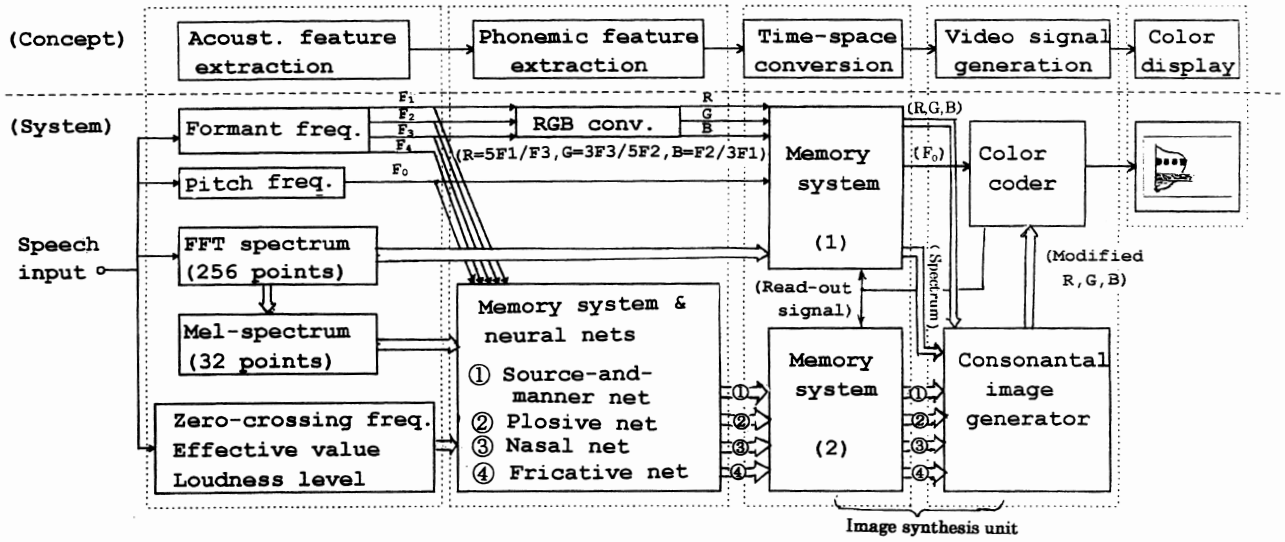


Fig. 1. Concept and system for speech-feature visualization.

We consider that the former example, in contrast with the latter case, shows special success in acquiring a scheme in the human brain for extracting phonemic features from speech spectrograms by means of extensive training. If this is true, the effective visual decoding of speech will be aided by technically realizing a feature-extracting scheme compatible with the one present within the auditory-neural processing of the human brain.

In general, normal hearing people unconsciously perceive various sound features in speech due to their mother tongue, such as distinctive features, prosodic features, phonemes etc. As a proof of it, they can pronounce almost any speech of their native language, and also understand or guess what speech conveys by those features, for example, meaning, emotion, voice quality, speaker's personality etc. Therefore, if such specific features of speech are skillfully integrated to form the visual image similar to the hearing image, they will contribute to the realization of the readable patterns of speech.

According to the above idea, we first developed the system to visualize continuous speech by integrating pitch and formant frequencies that were extracted in real time [10]. This system has some merit that can not be expected from other systems. For example, the color contrast is effective at restoring visually the deformation of acoustic parameters which the coarticulation effect causes, and as a result, it reduces the errors when reading three connected vowels [10], [11]. After developing this system, our effort for speech visualization has been directed toward realizing a true interface between speech features and visual sense in understanding speech.

From this point of view, the aim of this research was to complete the speech visualization system to allow us to intuitively read phoneme sequences of continuous speech after a comparatively short period of training. The proposed system and its performance will be described in this paper.

II. Summary of the Previous Systems [10], [12]

The speech visualization system, which we first developed with analog hardware, was reported in 1985 [10]. In the system, the formant frequencies that are extracted by inverse filtering, which we call the inverse-filter-control method [10], [13], are converted into three primary color signals by the following equations:

$$R = 5F_1/F_3,\ G = 3F_3/5F_2,\ B = F_2/3F_1. \qquad (1)$$

The circular ratios of the formant frequencies normalize the influence of vocal tract lengths on the formant frequencies, and the coefficients, 5, 3/5, and 1/3, make a neutral vowel colorless [10][14]. Thus, using (1), five Japanese vowels are represented as five different colors independently of the different vocal tract lengths due to the age and sex of the speakers.

The effect of the color representation in this system appears on connected vowels. That is, the visual tests verified that color contrast between adjacent vowels considerably compensated for the deformation of formant frequencies by coarticulation effects [10]. As results for reading three connected vowels, the correct answer rates were much higher for the proposed color patterns than for formant trajectories, which are similar to the dark lines in "visible speech." In these tests, two subjects tried to read three connected vowels (all combinations of /$V_1V_2V_1$/ syllables) uttered by two adult males and an adult female after learning the colors of the five isolated Japanese vowels by the same talkers. When the learning and test cycles were repeated, the learning curves for the subjects using color patterns were saturated at a score of 98% after 2–3 cycles, whereas the correct response rates corresponding to the formant trajectories barely reached the point of saturation at 72% after 6–7 cycles [10]. This effect was also verified by the tests reading a group of synthetic speech, which simulated three connected vowels. That is, the subjects responded to the stimuli as if the hue of the central (second) vowel had shifted in the direction of compensating for the coarticulation effect, owing to the color contrast that the first and third vowels had caused [11].

In 1988, we extended the system so as to represent consonant patterns by overlaying the sound spectrogram based on the Mel-band filter bank, on the color pattern [12]. After two subjects had memorized the vocabulary, which consisted of 50 nouns and 40 adjectives, they learned the corresponding patterns

uttered by a speaker using the extended system. When the other speakers uttered the same words in the test, the correct answer rate reached 85%. However, we felt intuitively that the sound spectrogram patterns could not present clear images peculiar to consonants. Provably, reading of the unknown words using this system would have been difficult. Based on the above results, we have searched for a new method to represent consonants and have applied it to the new system, which has been simulated by software of a personal computer. The new system does not operate in real time yet, but there is still a practical case for using the system for the hearing impaired as will be described in the conclusion. On the other hand, we are developing the real time system with digital signal processors (DSP) [15].

III. New Speech Feature Visualization System

Fig. 1 shows the processing concept and the concrete system for the new speech visualization, which has been proposed and realized in this research. Although the new system is based on the same concept as the previous ones [10], [12], the neural nets and the image synthesis unit play the most important roles in the new idea. An outline of the operation of the system is as follows: The first three formant frequencies, which have been extracted from speech signals frame by frame, are immediately converted into three primary colors (RGB signals) using (1) in Section II. The voiced-sound output from the neural net (① in Fig. 1) controls the magnitudes of the RGB signals with the consonantal image generator so that only the voiced parts are represented by colors. Pitch-frequency signals changing with time, which represent intonation, are converted into a change of horizontal length of the color pattern. FFT spectra with successive frames turn into sound spectrogram that shapes a white pattern by increasing the brightness. The phonemic features are extracted from the neural nets (①, ②, ③, ④) and generate the different consonant patterns as white textures with the consonantal image generator. The silence, which is extracted with the neural net (①) likewise, controls black so as to lower the brightness. The modified R, G, and B in Fig. 1 include all of the above-mentioned information as will be described in the Section (V). Finally, the color coder generates the composite color video signals from the modified R, G, and B. The memory systems [(1), (2)] are necessary to display time-varying features as a spatial pattern, which is easy to understand visually. The patterns flow from the bottom to the top of the screen so that we can always see patterns of continuous speech through a 2-s window. (The old pattern goes upward out of the top of the screen and the new one appears from the bottom, therefore, the time axis is directed downward.) If necessary, it is possible to stop the pattern on the screen. When several users want to use this system at the same time, the outputs of the color coder should be radiated from an antenna as weak electromagnetic waves. In this way, they can observe the patterns on several conventional TV sets.

Let us describe the main parts of our idea more in detail next. Vowels and semivowels are pronounced by a continuous change of place of articulation, while consonants are uttered from discrete places of articulation such as bilabial, alveolar, or velar position. Such properties in speech production influence the system components for extracting the acoustic and phonemic features in Fig. 1.

First, the continuous change of place-of-articulation is characterized by resonant (formant) frequencies in the vocal tract. So, we have used the improved inverse-filter-control method [17] for the formant extraction. Since describing the system in detail is difficult in this paper, let us summarize it shortly. In this method, many (16–32) inverse filters, each of which has two complex conjugate zeros in the system function, are mutually controlled by zero-crossing frequencies in their outputs. After the quick convergence of the inverse filters, speech signals are separated into a group of approximate single-resonant waves. Formant frequency is computed as the weighted mean in the zero-crossing frequency distribution of the approximate single-resonant wave. This principle is not directly dependent on spectral shapes to be influenced by bandwidths and amplitudes of the resonant components and moreover, doesn't need any criterion to minimize errors. As a result, the formant trajectories in this method are continuous and stable in time.

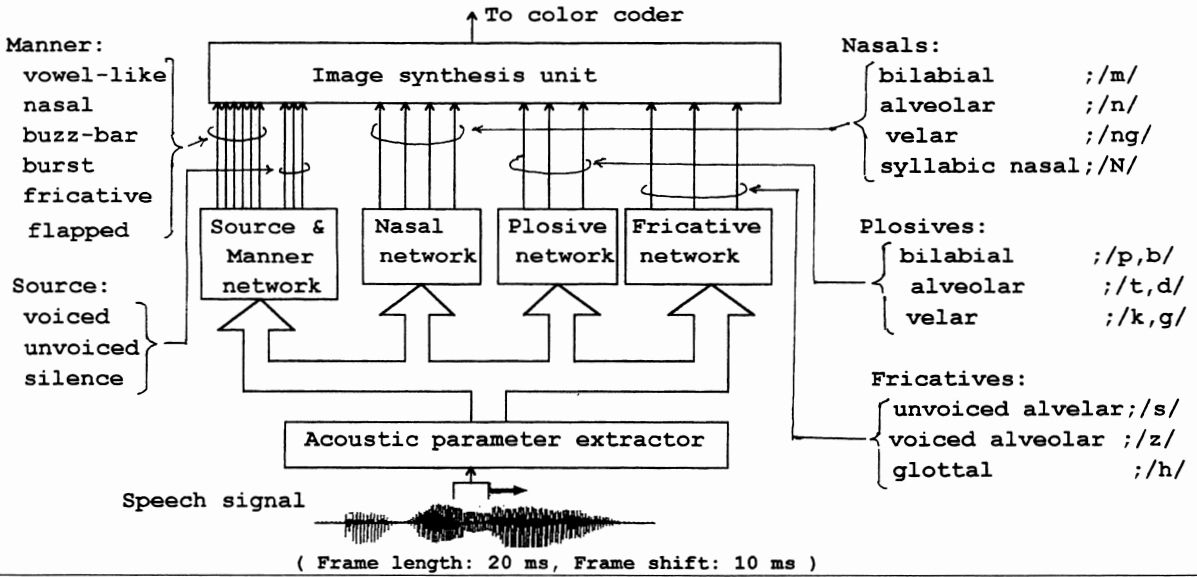
Second, the method for extracting phonemic features that represent consonants, is one of the main problems to be solved for this speech visualization. It is, in general, possible to indicate the acoustic features of consonants regarding perceptual cues that are determined by controlling synthetic speech, like F_2 locus in the voiced plosives /b, d, g/ [16]. However, we often experience that the consonantal features in real speech are analytically obscure despite the perceptually clear image of the phonemes.

A hypothesis to explain the above facts is that some features, each of which is too weak individually to characterize the phoneme, support the heard perception of the phoneme by helping one another. Accordingly, except a few consonants, we don't have unfortunately any excellent and explicable algorithm for the consonant-feature extraction although we can understand some perceptual cues and the models for generating consonants. In this case, a neural network will be useful because it can learn the relative differences between compound properties in various consonants and separate one phonemic feature from the others. We have four neural networks to extract the phonemic features of consonants for the new system. The construction and function of the neural nets will be described in the next section.

Last, the sound spectrograms are overlaid on colors in voiced parts so that those looks white for the increased brightness. We expect, in this system, that sound spectrograms, which FFT spectra make with a linear frequency scale, are effective to get rough information about speakers rather than phonemes in voiced sounds, because phonemes can be represented by the features other than FFT spectra. When the speaker is a child with high pitch, the spectrograms of voiced sounds represent coarse and waving stripes of harmonic components and it impresses child's voice intuitively. The represented spectrograms of voiced sounds have been restricted below 2.5 kHz not to disturb the color pattern of voiced parts though the unvoiced parts have been represented in a full range.

IV. Phonemic Feature Extraction Using Neural Networks

In general, phonemes are classified using three parameters (distinctive features) for speech production, that is, a sound source, manner-of-articulation and place-of-articulation. Since



[Acoustic parameters for neural network inputs]	
(a) 32 channel Mel-spectrum	(e) Effective level
(b) The lowest four formant frequencies (F1-F4)	(f) Loudness
(c) Pitch frequency (F0)	(g) Existence or non-existence of F0
(d) Zero crossing frequency	(h) Existence or non-existence of formants

Fig. 2. Neural networks for extracting phonemic features.

each parameter has discrete categories to classify phonemes, neural networks with a supervisor will be available to the extraction of phonemic features.

Thus, to specify the phonemic features presenting in speech signals, we used four neural networks, which are three-layer perceptrons with the following output categories.

1) Source and manner-of-articulation network (neural net for sound source and manner-of-articulation).

   This neural net is trained to classify the sound source of input signals into any one of voiced, unvoiced and silence as output categories.

   The manner-of-articulation categories are defined as vowel-like, nasal, buzz-bar, plosive, fricative and flapped (/r/).

2) Plosive network (neural net for detecting place-of-articulation in plosives).

   Bilabial (/p/ or /b/), alveolar (/t/ or /d/) and velar (/k/ or /g/) are necessary for the output categories.

3) Nasal network (neural net for detecting place-of-articulation in nasals).

   Bilabial (/m/), alveolar (/n/), velar (/ng/) and syllabic nasal (/N/) constitute the output categories.

4) Fricative network (neural net for discriminating three kinds of fricatives).

   Unvoiced alveolar (/s/), voiced alveolar (/z/) and glottal (/h/) are extracted by this neural net.

Fig. 2 shows the system construction with the neural nets and the acoustic parameters as inputs. The parameters are of eight kinds, which contain 42 acoustic features in all. Speech waves have been sampled at 12 kHz and the acoustic features for the neural net inputs have been extracted every single frame. The frame length and the frame shift are 20 ms and 10 ms respectively. The outputs of the neural nets become the inputs of the image synthesis unit, which consists of the memory system (2) and the consonantal image generator shown in Fig. 1.

The preliminary experiments to determine appropriate conditions for the above neural nets showed a need for 30 units in the hidden layer, and in addition, the method chosen to accelerate convergence in the back-propagation method [18], was set the momentum to 0.8 and the learning rate to 0.2.

In many experiments using time-delay neural networks (TDNN's) for speech processing, many successive frames of a single parameter have been used as inputs [19], [20]. However, the use of different kinds of acoustic parameters, which represent various attributes of speech, may be more useful than a single one.

According to our previous experiments, eight kinds of acoustic parameters, which consist of 42 features in all, were effective in extracting the phonemic features by the perceptron even if features of only isolated frames were used [21].

In this research, for the same eight kinds of parameters as above, we have tried to get better results using three selected frames, which are not necessarily successive. This means we extend the perceptron into another type of TDNN.

The use of three frames requires 126 parallel inputs for each perceptron. Next, we searched for three frames that provide a high recognition rate for each neural net. We have estimated the recognition rate by judging if the maximum-output category is correct or not, when compared with the phonemic label due to visual inspection of speech waves and spectrograms. The search method we used to select the effective frames is as follows.

The training materials for the neural nets are 62 /VCV/ syllables, which have been uttered by 20 males

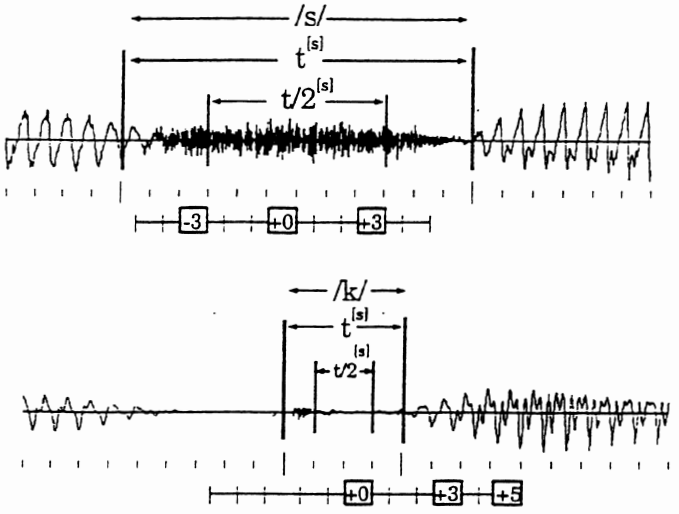


Fig. 3. Examples of selecting input frames from consonantal parts: The standard frame (+0) is randomly selected from the central part and combined with the other frames (+3, −3 etc.).

TABLE I
COMBINATION OF INPUT FRAMES TO BE USED FOR THE NEURAL NETS

Name of neural network	Selected frames
Source and manner	0, +3, -3
Nasal	0, +5, -5
Plosive	0, +3, +5
Fricative	0, +3, -3

and 20 females. The test materials consist of the same syllables as those for the training, but uttered by another ten males and ten females.

First of all, a standard frame has been chosen randomly from the central part, which means 50% of time length in the consonantal part as shown in two examples of Fig. 3. Next, we have decided pairs of frames that give the highest recognition rate, through the training and test for selecting another frame combined with the standard one. Finally, three frame combinations have been decided likewise by adding a third frame to the previously fixed frames.

Since it is desirable that the selected frames are common to four neural nets, we have selected four sets of three frames, as common as possible, from several combinations that give the recognition rate near the maximum.

The frame combinations obtained by the above approach are shown in Table I. The standard frame "0" indicates a randomly selected frame from the central part of each phoneme segment as indicated in Fig. 3. The frame "$+m$" indicates a frame whose distance from the standard is $+m$. In the source-and-manner net, the nasal net and the fricative net, the additive frames have been selected symmetrically, i.e., $\pm m$. This symmetry probably depends on a quasistationary property of their features. In contrast, the "0," "+3" and "+5" frames have been chosen for the plosive network. It seems to indicate that the place-of-articulation information in plosives is contained around the beginning of vowel closely following the burst.

The results of the final tests using the selected frames are shown in comparing with them of the isolated frames in Table II. In nearly all cases, the recognition rates greatly increase by adopting the features of three frames. This has been confirmed

TABLE II
INCREASE OF THE RECOGNITION RATES WHEN USING THREE INPUT FRAMES

(a) Source and manner-of-articulation network

	Male speakers		Female speakers	
	Number of frames			
< Categories>	1	3	1	3
Vowel-like	86.1 (%)	92.1 ** (%)	83.3 (%)	87.9** (%)
Nasals	65.9	74.0 *	61.6	65.9 *
Buzz-bar	74.1	81.9 *	69.7	76.4 *
Voiced plosives	71.4	90.7 ***	71.5	77.8 *
Unvoiced plosives	65.9	73.6 *	69.1	84.6 ***
Voiced fricatives	86.4	91.3 *	55.0	68.8 ***
Unvoiced fricatives	89.5	92.1 *	91.1	94.8 *
Flapped /r/	75.2	86.5 ***	65.7	80.0 ***
Silence	96.9	99.5 *	98.8	99.6 *
Mean	79.0	86.8 ***	74.0	81.7 ***

*Significant at the level p<0.05, **p<0.01, *** P<0.001

(b) Plosive network

	Male speakers		Female speakers	
	Number of frames			
<Categories>	1	3	1	3
/p/, /b/	84.4 (%)	87.1 (%)	83.8 (%)	90.5 * (%)
/t/, /d/	80.5	86.3	73.7	87.7 *
/k/, /g/	90.4	93.4 *	87.3	89.9
Mean	85.1	88.9 *	81.6	89.4 *

* Significant at the level p<0.05

(c) Nasal network

	Male speakers		Female speakers	
	Number of frames			
<Categories>	1	3	1	3
/m/	49.9 (%)	72.5 * (%)	33.0 (%)	65.3 *** (%)
/n/	36.5	58.5 ***	45.4	60.6 ***
/ng/	45.3	54.4 *	44.7	53.8 **
/N/	64.1	94.2 **	69.3	94.0 ***
Mean	48.9	69.9 ***	48.1	68.4 ***

*Significant at the level p<0.05, ** p<0.01, ***P<0.001

(d) Fricative network

	Male speakers		Female speakers	
	Number of frames			
<Categories>	1	3	1	3
/s/	96.0 (%)	97.5 (%)	88.2 (%)	94.5 ** (%)
/z/	90.1	95.6 *	79.8	88.3 **
/h/	93.1	93.0	94.9	93.7
Mean	93.3	95.4 **	87.6	92.17 **

*Significant at the level p<0.05, ** p<0.01

using a test of significance, i.e., A student's t-test for differences between two averages.

V. IMAGE SYNTHESIS TECHNIQUE FOR CONSONANTS

The image synthesis technique we used for the representation of consonants is based on an essential idea for generating the image naturally. Fig. 4 shows the processing flow for the image synthesis unit of Fig. 1. First of all, we prepared textures peculiar to each category of the manner-of-articulation as bitmap patterns for a whole area of the screen. Those textures are necessary to represent visually the auditory image of the manner-of-articulation. For example, we intuitively represented nasal sounds by a mesh, plosives by horizontal bars, fricatives by random or arranged small dots etc. The display position at which the visual pattern is put connects roughly the place-of-articulation in a vocal tract. That is, bilabial, alveolar and velar consonants correspond to right, middle and left positions on the screen respectively. The pattern images (textures) and the display positions are depicted as caricatures for every consonant in the middle of Fig. 4.

Next, a set of neural-net outputs is read out from the memory system (2) by the horizontal hold control signals of the color coder, and they control the brightness of the visual patterns for

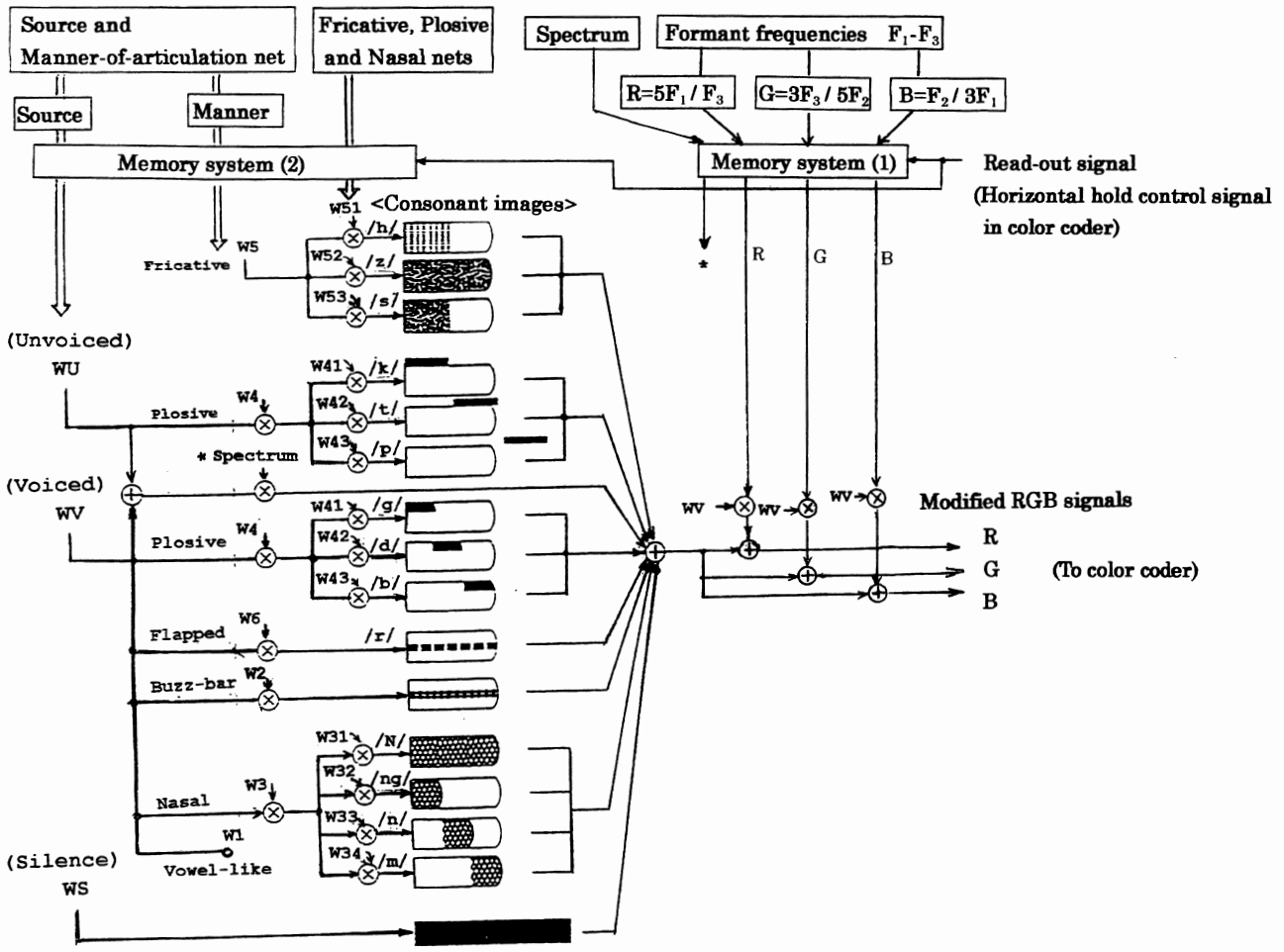


Fig. 4. Processing flow in image synthesis unit.

one scanning line on the screen as follows. As described in the previous section, a phoneme is specified by outputs from three sets of parameters that represent sound source, manner- and place-of-articulation. Therefore, a product of those outputs can be used as a score for how much the phoneme appears. (Fricatives /*s*, *z*, *h*/ and flapped /*r*/ need the product of two outputs only, as shown in Fig. 4.) Multiplying each score by the corresponding bitmap pattern determines the relative brightness of the phonemic pattern. For example, as the score of consonant /*m*/, the product of three outputs, voiced in "sound source," nasal in "manner" and bilabial in "place," is computed ($WV \times W3 \times W34$) and then it is multiplied by the bitmap pattern of /*m*/. As a result, the brightness of visual pattern, /*m*/, is determined for one scanning line.

The brightness of each phonemic pattern, in this manner, is computed individually. As an example, the brightness signals before multiplying by the bitmap pattern, i.e., the excitation signals for the phonemic (bitmap) pattern, are shown together with the formant and pitch frequency trajectories in Fig. 5.

After adding all the phonemic patterns weighted by the brightness, the resultant signal is added to each of the three primary color signals. Finally, those modified signals of three primary colors are transmitted to the color coder, which generates the composite color video signal. When these operations are repeated for all of the scanning lines of a whole screen, we get a spatially changing visual image of speech signals lasting up to 2 s.

This technique automatically depicts the phonemic sequence with high brightness on the screen. Possibly two kinds of phonemes with nearly equal brightness appear at the same time in a word, but it is basically allowable, or rather natural because we sometimes receive the ambiguous sound for a phoneme in auditory perception too. The redundancy of context will lead us to correct judgment in such cases. This is, we consider, one of the advantages for the context-free visualized speech that is different from alphabetic notation to be determined by speech recognition.

The speech image flows from the bottom to the top on the screen in real time if the contents of the memory systems (1) and (2) are continually updated. In this way, we always see continuous speech through a 2-s window.

VI. Display Examples of Japanese Words, Comparison with Speech Spectrograms

In order to show intuitively what effective information the system presents, four color pictures of the visualized speech uttered by males are shown in comparison with spectrograms in Figs. 6–9. The speech spectrograms were made using SP4WIN Pro. (Version 1.0) developed by NTT-AT Co. Ltd. In the color

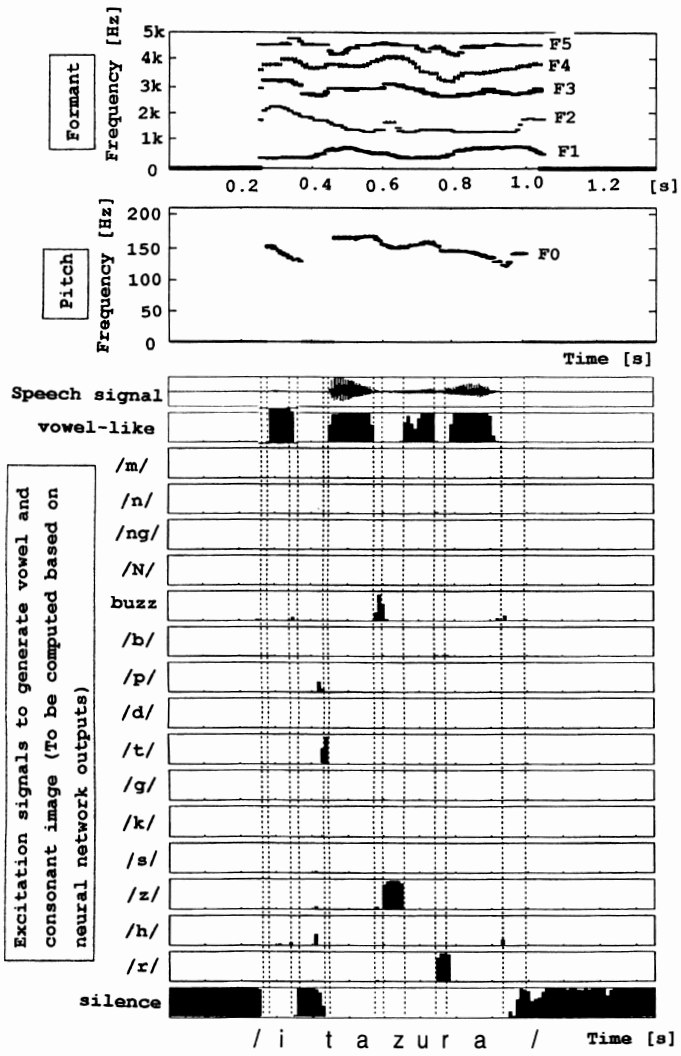


Fig. 5. Examples of phonemic features to create a picture (Japanese word, /itazura/).

pictures, the vertical axis represents time and the horizontal length of the color pattern indicates pitch. The pitch changes with time, i.e., intonation appears clearly in all of the pictures. The color image corresponds to vowel quality. The pattern image (texture) and its position indicate what the consonant is, as illustrated at the center of Fig. 4. The brightness difference between colors doesn't appear because of a luminance correction [10], so we see vivid colors over the whole of the screen in Figs. 6–9.

The picture of Fig. 6 has been created from the data shown in Fig. 5. As shown in Fig. 5, the adopted formant detector extracts formant frequencies even in very low level portions of speech signals. On the other hand, the brightness of phonemes (the excitation signals in Fig. 5), which has been determined by the outputs of the neural nets, indicates the extent of appearances of the vowel-like sound, each of the consonants and the silence. According to the respective brightness, the transparent pattern for the vowel-like, the white textures for the consonants and the black for the silence are overlaid on the colors due to formant frequencies. By this processing, the black in the closure period just before plosives and the other visual features are represented manifestly as in the picture of Fig. 6. Thus, in the adopted method, the system integrates the phonemic features naturally and doesn't use any technique for segmentation and recognition.

From the pictures in Figs. 6–9, we observe the following characteristics. There is almost no perceptual error in the words /itazura/ ("mischief" in English) and /subarashi:/ ("wonderful"). On the other hand, /sekiraNuN/ ("cumulonimbus") will be corrected by redundancy of the meaningful word from the meaningless word /sekiraNmuN/ as a phoneme sequence. Moreover, we observe the overlapping representation of two kinds of patterns, that is, /m/ itself and the buzz-bar in /m/ of /puroguramu/ (Japanese pronunciation of "program"). In this case also, the buzz-bar will be neglected owing to the redundancy of the meaningful word.

Comparing the four pictures created by our system, we find slight differences among the colors of the same vowel in the different words. These differences are mainly due to the speaker's individuality and/or coarticulation effects in context. If we perceive those colors as belonging to the same vowel category, then we will be able to read correctly Japanese words. English vowels have more than five categories. So, colors with slight differences may be perceived as belonging to different categories.

When we see an F_1-F_2 diagram of isolated vowels uttered by many talkers, we notice the following characteristics. A specific Japanese vowel shows relatively large dispersion [10], [22], while the area of one specific vowel may be separated into 2–5 different English vowels, with relatively small dispersion in each vowel [23]. (English vowel, /æ/ rarely appear in isolated Japanese vowels.) Therefore, we expect that learning of the correspondence between auditory and visual images will foster perceptual constancy of the visual image in different languages.

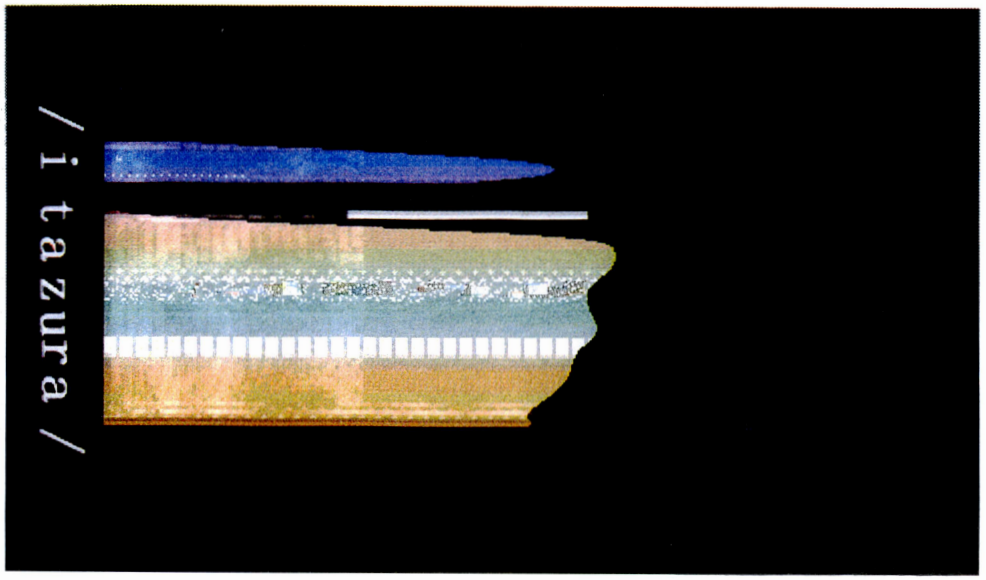
The strength of this representation is in that we can indicate visually the clear phonemic sequences and boundaries. In addition, the color of the vowel sandwiched between consonant patterns looks uniform and semivowels are characterized by a change from an intermediate color to a target one. These manifest representations enable us to easily read the phoneme sequences uttered by the different speakers.

In the case of speech spectrograms, it is often hard to correctly read phonemes apart from the context because the formant movement plays an important role in the visual segmentation. It will be difficult in particular for ordinary readers to correctly read vowels and some consonants which are influenced by the individual differences of speakers, unless relatively long utterances are represented on a sheet of paper or a screen.

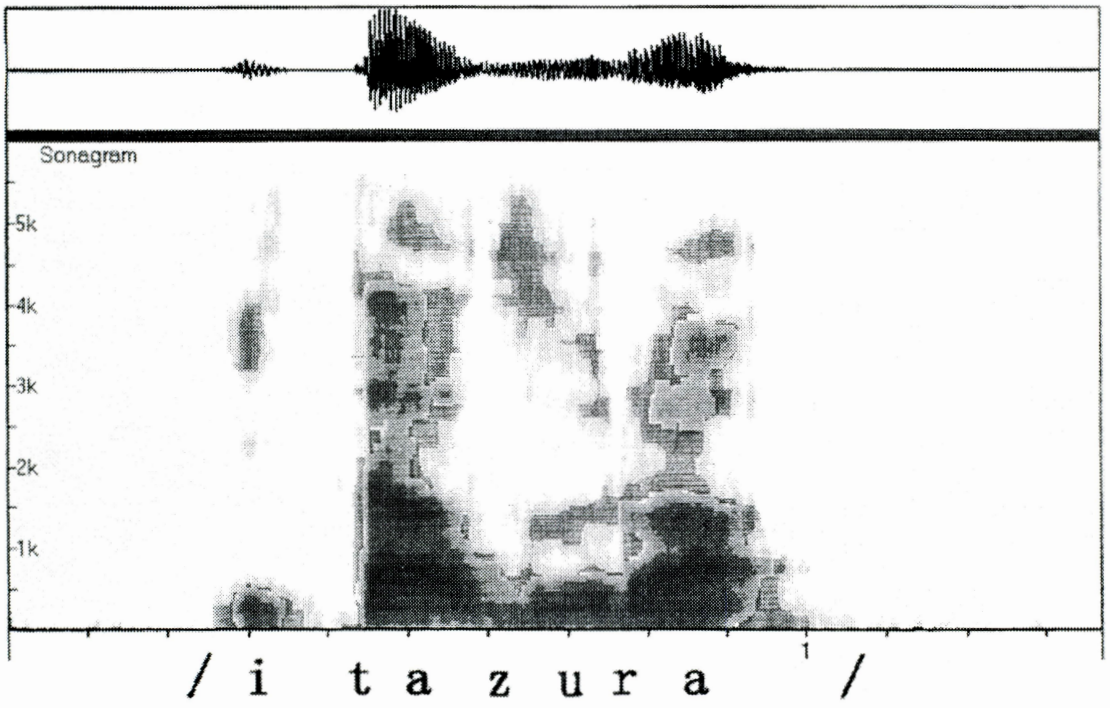
If we read both representations through a 100 ms short-term window (slit), then it will be clear that the representation proposed in this paper is far less dependent on the context and the individual difference of utterances than the speech spectrograms are. The above differences between the proposed visualization and the speech spectrograms will influence the intuitive readability.

VII. Preliminary Test for Reading Words Using the Proposed System

The first test using the proposed system was carried out to investigate readability of the realized representations and durability of the learning effect. For this purpose, we used household words uttered by adult males as speech materials for the tests.



(a) Visualized speech (proposed)



(b) Speech spectrogram

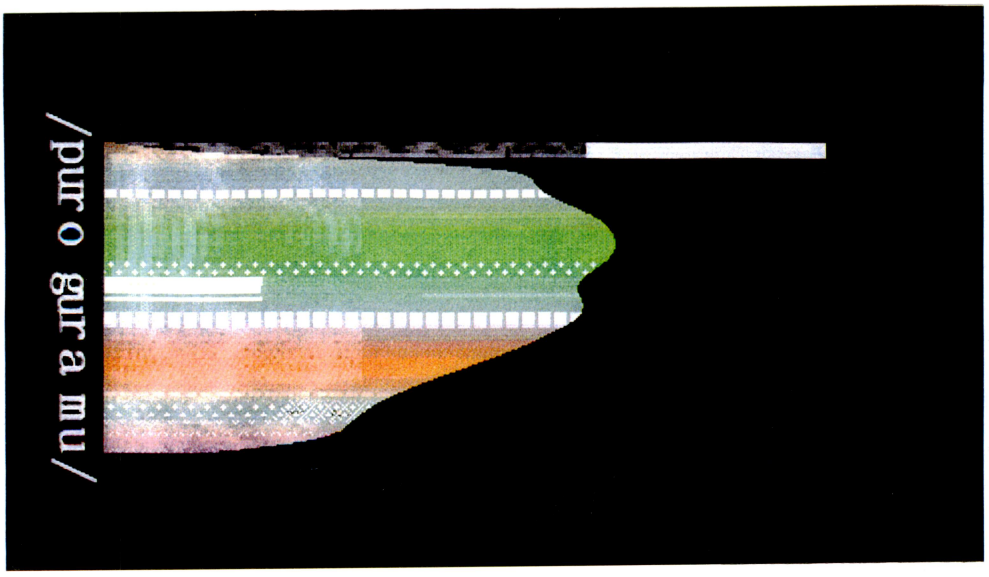
Fig. 6. Display example of visualized speech in comparison with speech spectrogram [Japanese word, /itazura/ ("mischief")].

A. Subjects

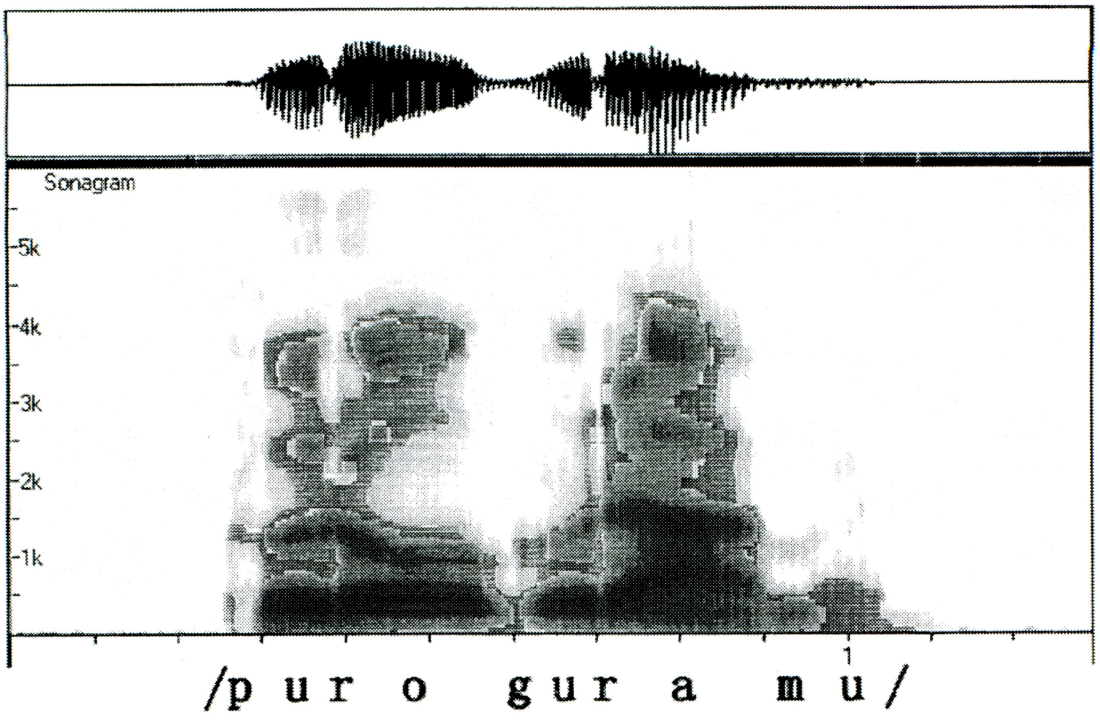
Three students with normal hearing took part in reading tests as subjects. All of them were graduates or undergraduates engaged in the study of speech information processing. Although they understood roughly how to create the visual patterns using the speech parameters before the reading tests, none of these students had extensive knowledge of phonetics.

B. Method and Results

The subjects need training to understand intuitively the visualized speech. The training has been performed before the real



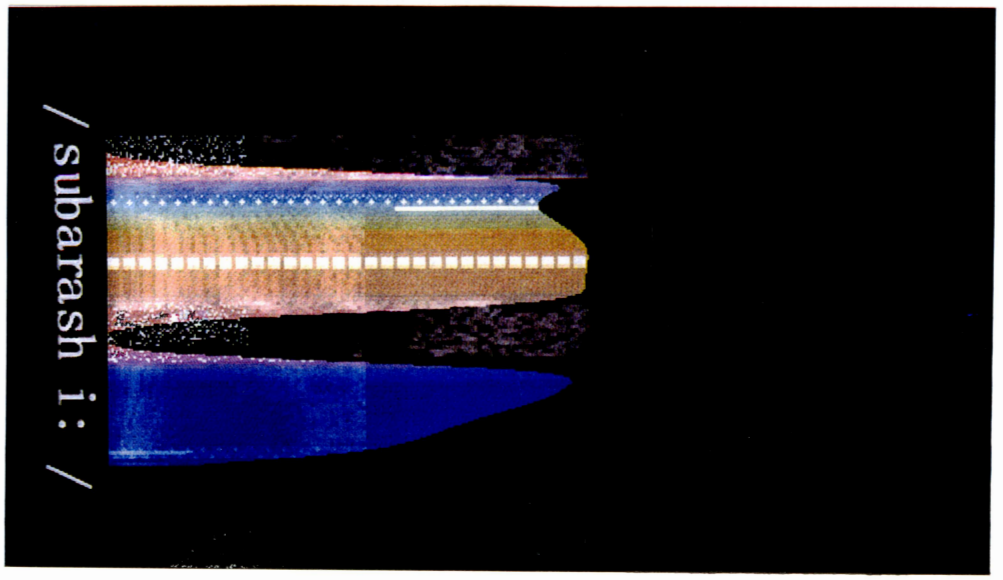
(a) Visualized speech (proposed)



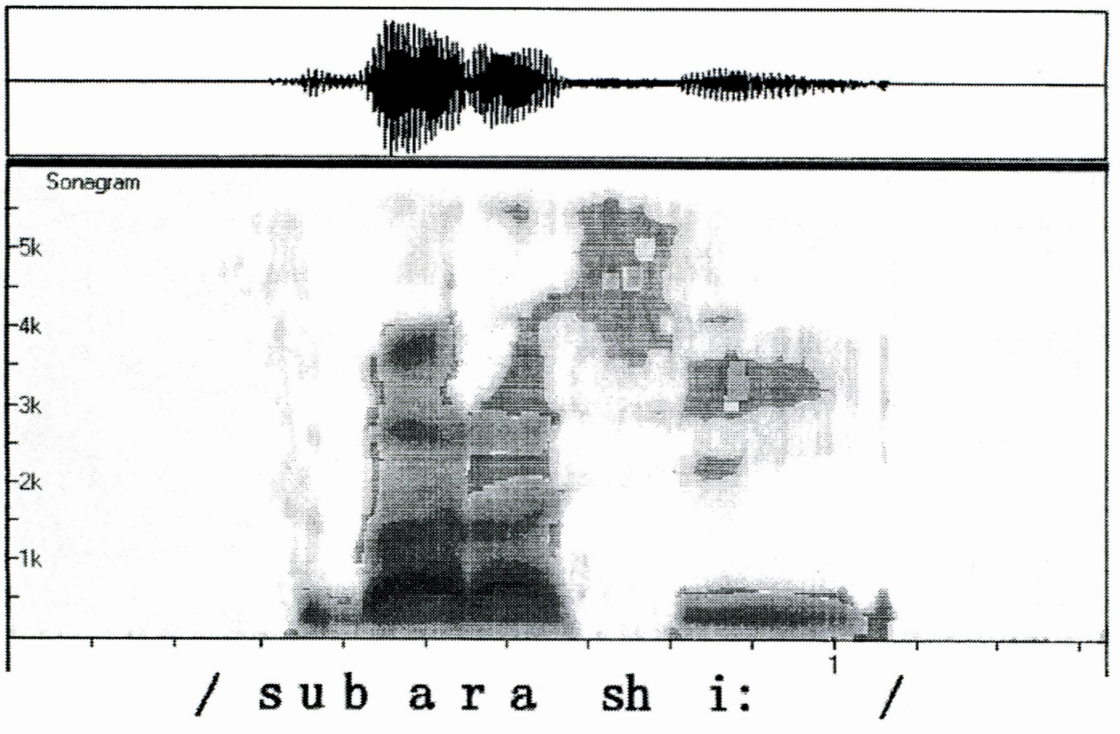
(b) Speech spectrogram

Fig. 7. Display example of visualized speech in comparison with speech spectrogram [Japanese word /puroguramu/ ("program")].

tests to estimate the subjects' reading ability. The training stage consists of a learning stage and a confirming test. In the learning stage, a pair of speech sound and its visual pattern has been presented at the same time, one by one. Next, in the confirming test, the learning effect was confirmed by a simple test using a subset of the visual patterns only. The speech materials prepared for the learning stage consist of 62 /VCV/ syllables, 101 monosyllables (CV) and 50 meaningless words. /VCV/ and /CV/ syllables pronounced by a male have been processed to make all of the patterns errorless for efficient training. That is, when unclear consonants were found in the syllables, we replaced the consonant parts by the artificial patterns, which were correct and clear,



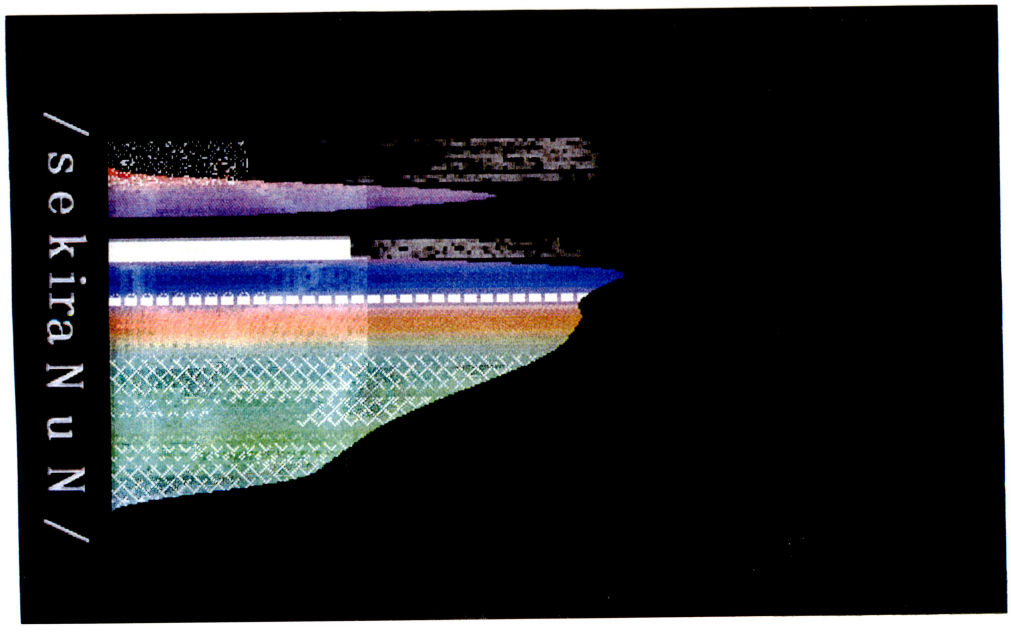
(a) Visualized speech (proposed)



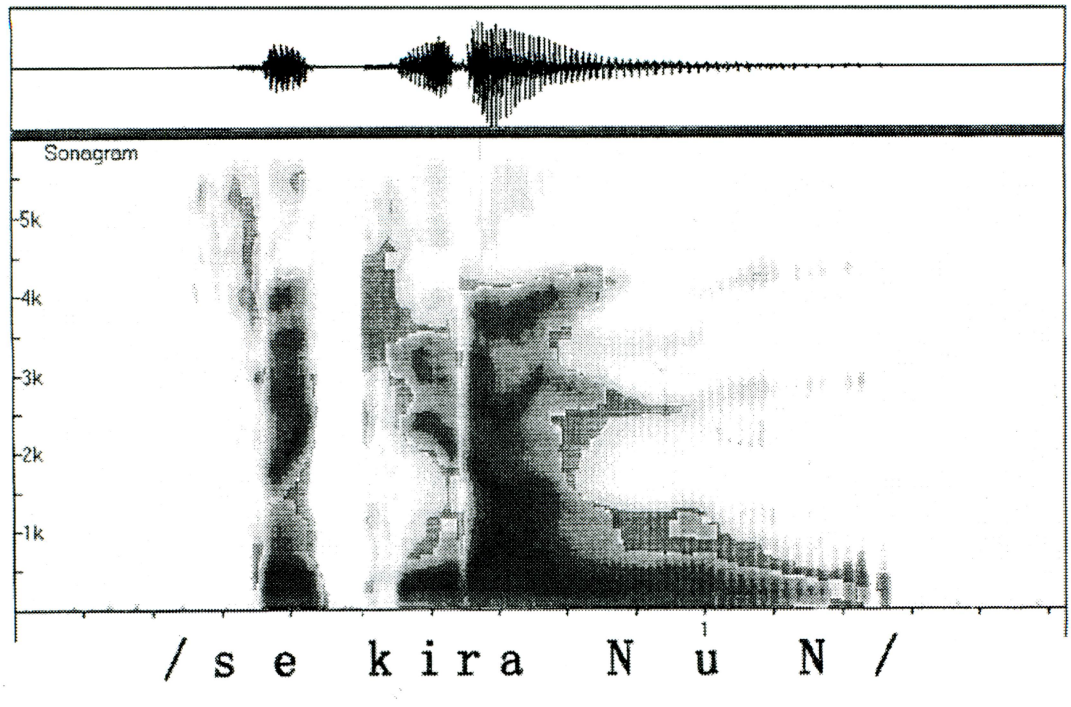
(b) Speech spectrogram

Fig. 8. Display example of visualized speech in comparison with speech spectrogram [Japanese word /subarasi:/ ("wonderful")].

leaving the vowels unchanged. The meaningless words, which have been uttered by two males, are almost phonetically balanced and each of them is constructed with 2–4 syllables. From the 50 meaningless words, 25 words have been extracted randomly every one learning stage and used each time, but, /VCV/ and /CV/ syllables have been presented at the first several times only, and after that, according to the subject's request. In the confirming test, we prepared 75 meaningful words uttered by the same talkers; 25 meaningful words of them were used first as a list. After the correct response rate to be obtained by the confirming test had reached a sufficiently high score (more than 90%), 25 new meaningful words were appended to the list and



(a) Visualized speech (proposed)



(b) Speech spectrogram

Fig. 9. Display example of visualized speech in comparison with speech spectrogram [Japanese word /sekiraNuN/, ("cumulonimbus")].

the training continued likewise. Finally, when the confirming test using 75 meaningful words by appending the new other words showed good results, the training session was closed and the real test session started. The number of the trials necessary to reach the point of saturation in the correct response rate was at most 8–10 times in any of the three confirming tests. Once the real test had begun, the training was never done again. All the patterns were displayed in stop motion in the center of the screen. In the tests, the subjects were required to push a button immediately once they had the answer. The visual pattern then disappeared and the subjects answered orally toward the recording system. The time interval between the instants

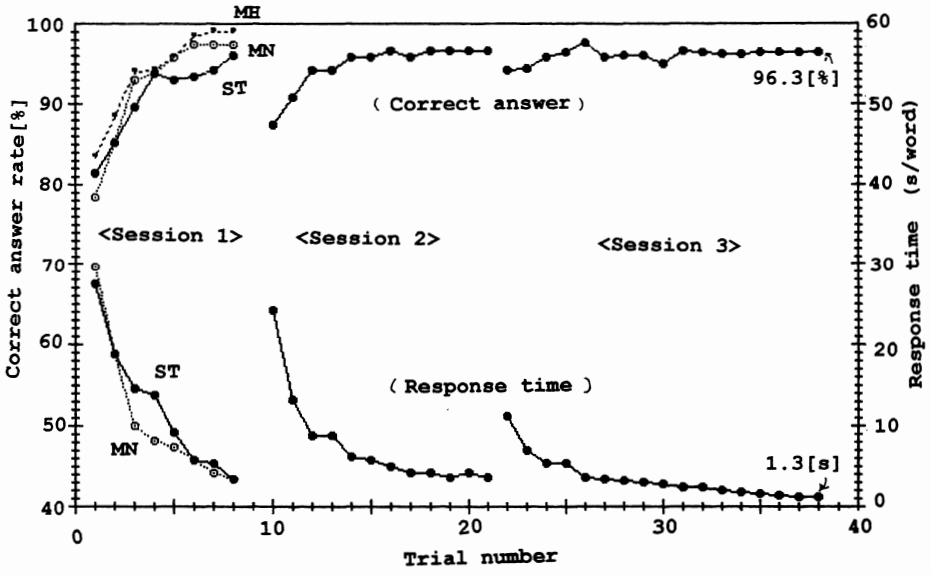


Fig. 10. Learning curves in preliminary test (reading test for 300 tokens, which consist of 75 words uttered by four males).

of the pattern presentation and pushing the button gives the response time. During all the tests including both confirming and real tests, the subjects were not told correct answers. However, they will have guessed the vocabulary based on their own answers by repetition of the test. This means that the test itself causes the subjects effects equal to learning.

In the real test sessions, four male talkers uttered 75 new meaningful words, which had never been presented in the confirming tests. These Japanese words consist of 15 two-syllable words (/inu/, /ushi/, /kaze/ etc.), 30 three-syllable words (/kitsune/, /hagaki/, /baiku/ etc.) and 30 four-syllable words (/shimauma/, /koueN/, /keshigomu/ etc.). Since the talkers tried to pronounce naturally, partial deformations like devocalization of vowels (/sh(i)mauma/, /kesh(i)gomu/ etc.) were also contained in those words. All of the words (300 items) were randomized and used for every trial. After the subject had answered a pattern in the test, the experimenter presented the next word by manual operation of the computer. Therefore, though the time necessary for one trial of the real test was very long (about 2.5 full hours) at first, it was drastically getting shorter as the test proceeded. At the final trial, the test ended within 15–20 min. In the case of a trial lasting beyond 30 min, we divided the trial into units of about 30 min, for example, by repeating 30 min of test and 15 min of rest. The real tests were carried out as three sessions shown in Fig. 10. The period per session was 30–40 days and about five months of rest were put between the sessions. To examine durability of learning, we didn't at all carry out retraining before or during the last two sessions.

As shown in the results of Fig. 10, the learning curves of three subjects in the first session almost overlap and show a steep rise. Moreover, their correct answer rates reach 96–99% after eight trials. The subject ST maintains his reading ability considerably in the succeeding sessions despite inactive periods of five months. Finally, his reading ability attains an understanding of 96.3% of 300 tokens and a response time of 1.3 s per word after the reading experience of about 40 full hours in all including the training and tests. As the response time includes the time necessary to push a button, it shows that he can read the visual patterns almost instantly.

VIII. Conclusion

We have completed a speech visualization system based on the new application of extracting the consonantal features using neural networks (TDNN's). In the proposed system, four kinds different neural nets determine the strengths of the features for sound source, manner- and place-of-articulation. For high performance, we selected three input frames of the acoustic parameters. The effect of using three input frames, which are not immediately adjacent, was statistically significant in nearly all cases, when compared with using just one isolated frame.

Next, we have developed a technique for creating visual images by simply adding all the consonantal patterns whose brightness is controlled by the strength of the extracted phonemic features. Some examples of the visualized speech indicate that the visual images of consonants make the phonemic sequences and boundaries clear, and that the color of a single vowel sandwiched between consonants looks uniform. These representations, which need neither segmentation nor recognition, are effective for the readers to understand speech intuitively.

Finally, we have evaluated the performance of the new system in a preliminary test in which three students read the visual patterns of 75 words uttered by four males (300 items). The learning curves showed a steep rise and attained 96–99% after eight trials. The response time was getting much shorter as the test proceeded and reached 1.3 s/word at the final trial. The learning effect was durable despite long rests.

According to visual inspection and reading tests of the pictures, the effect of the proposed representation on visual perception of speech is to give us a visual image which can be understood as easily as heard speech. So, we have a hypothesis that the signal processing and the parameter integration used in this research possibly exist in some equivalent form in auditory nervous system. Anyhow, since integrating or adding simply the phonemic features converts speech signals into recognizable

segments, we conclude that the process of phonemic feature extraction and image synthesis essentially makes visual decoding of speech possible. In other words, we believe that it will be difficult to acquire the visual decoding of speech by the unintegrated representation of acoustic features only.

The proposed system will be useful for new applications as well as for conventional ones such as speech training and speech transmission. For example, we notice that CD-ROM's, in which speech sounds and those pictures are recorded, will be useful for personal auditory training using a conventional PC and a hearing aid [24], [25]. Specifically, this system will be used to enable hearing-impaired persons to clearly perceive the subtle cues of speech sounds, which are provided by a hearing aid or a cochlea implant system, for example.

The success of visual decoding also seems to be related to the fact that these phonemic features are very effective for the speaker-independent word recognition, in which the processing algorithm for compensating coarticulation effects is not especially considered [26], [27].

ACKNOWLEDGMENT

The authors would like to dedicate this paper to the late S. Kisu who put his heart into the early phases of this work. They also wish to thank deeply Dr. Y. Ueda, T. Ikeda, N. Ikeda, and K. Iwata for their valuable comments and collaborations. They also wish to thank many graduates and undergraduates who collaborated in this research. They especially appreciate the collaboration conducted by Y. Nishida, S. Tokunaga, Y. Fukushige, and H. Ling.

REFERENCES

[1] R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech*. New York: Van Nostrand, 1947.

[2] R. Biddulph, "Short-term autocorrelation analysis and correlatograms of spoken digits," *J. Acoust. Soc. Amer.*, vol. 26, pp. 539–541, 1954.

[3] S. H. Chang, G. E. Pihl, and J. Wiren, "The intervalgram as a visual representation of speech sounds," *J. Acoust. Soc. Amer.*, vol. 23, pp. 675–679, 1951.

[4] P. A. Mitchell and R. D. Easton, "Wave collation visual speech display: Design and evaluation," *J. Acoust. Soc. Amer.*, vol. 97, pp. 1297–1306, 1995.

[5] G. M. Kuhn, "Description of a color spectrogram," *J. Acoust. Soc. Amer.*, vol. 76, pp. 682–685, 1984.

[6] L. C. Stewart, W. D. Larkin, and R. A. Houde, "A real time sound spectrograph with implications for speech training for the deaf," in *Proc. ICASSP 76*, 1976, pp. 590–592.

[7] R. A. Cole, A. I. Rudnickey, V. W. Zue, and D. R. Reddy, "Speech as patterns on paper," in *Perception and Production of Fluent Speech*, R. A. Cole, Ed. Hillsdale, NJ: Lawrence Erlbaum, 1980, pp. 3–50.

[8] B. G. Greene, D. B. Pisoni, and T. D. Carrel, "Recognition of speech spectrograms," *J. Acoust. Soc. Amer.*, vol. 76, pp. 32–43, 1984.

[9] A. L. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Why are speech spectrograms hard to read?," *Amer. Ann. Deaf*, vol. 113, pp. 127–133, 1968.

[10] A. Watanabe, Y. Ueda, and A. Shigenaga, "Color displaysystem for connected speech to be used for the hearing impaired," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 164–173, 1985.

[11] Y. Ueda and A. Watanabe, "Visible/tactile vowel Information to be transmitted to the hearing impaired," *J. Acoust. Soc. Jpn.*, vol. 8, pp. 99–108, 1987.

[12] A. Watanabe and Y. Ueda, "Speech visualization and its application for the hearing impaired," *J. Acoust. Soc. Amer.*, vol. 84, p. 42, 1988.

[13] A. Watanabe, "A real-time formant tracker using inverse filters," Tech. Rep. STL-QPSR 3-4, pp. 1–30, 1979.

[14] J. M. Pickett, *The Sounds of Speech Communication*. Baltimore, MD: University Park Press, 1980, pp. 46–49.

[15] T. Ikeda and A. Watanabe, "A DSP system for real-time speech processing and its application to parameter estimation," in *Proc. 8th Digital Signal Processing Symp.*, 1993, pp. 321–327.

[16] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries," *J. Exper. Psychol.*, vol. 54, pp. 358–368, 1957.

[17] A. Watanabe, "Formant estimation method using inverse filter control and its application to speech visualization" (in Japanese), in *Proc. Autumn Meeting Acoust. Soc. Jpn., (Special Session)*, 1998, pp. 521–524x.

[18] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE Acoust., Speech, Signal Process. Mag.*, pp. 4–22, Apr. 1987.

[19] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural network," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 328–339, 1989.

[20] J. B. Hampshire II and A. H. Waibel, "A novel objective function for improved phoneme recognition using time-delay neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp. 216–228, 1990.

[21] A. Watanabe, S. Tomishige, and S. Tokunaga, "Speech visualization by extracting features with neural networks," in *Proc. 15th ICA*, vol. 3, 1995, pp. 185–188.

[22] S. Saito and K. Nakata, *Fundamentals of Speech Information Processing* (in Japanese). Tokyo, Japan: Ohm, 1981, p. 37.

[23] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175–194, 1952.

[24] T. Ikeda, A. Watanabe, M. Hino, and Y. Ueda, "A training system to improve the usefulness of hearing aids," in *Proc. 16th ICA/135th ASA*, 1998, pp. 219–220.

[25] T. Ikeda, A. Watanabe, K. Yamazoe, and Y. Ueda, "An audio-visual system for telephonic speech to improve hearing ability of the hearing impaired," in *Proc. ASA-ASJ 3rd Joint Meeting*, 1996, pp. 273–276.

[26] T. Dutono, N. Ikeda, and A. Watanabe, "Effects of compound parameters on speaker-independent word recognition," *J. Acoust. Soc. Jpn.*, vol. 19, pp. 1–11, 1998.

[27] ——, "Word recognition experiments for evaluating compound features of speech," *J. Acoust. Soc. Jpn.*, vol. 19, pp. 155–157, 1998.

Akira Watanabe received the B.E. degree in electrical engineering in 1962, and the M.E. and D.E. degrees in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1964 and 1968, respectively.

He has been with the Faculty of Engineering, Kumamoto University, Kumamoto, Japan, since 1967. He investigated a real-time formant estimation system from 1978 to 1979 as a Guest Researcher at the Royal Institute of Technology (KTH), Stockholm. He is now a Professor with the Department of Computer Science, Kumamoto University. His research interests include speech processing and coding for the hearing impaired.

Shingo Tomishige received the B.E. and M.E. degrees in electrical engineering and computer science, Kumamoto University, Kumamoto, Japan, in 1994 and in 1996, respectively, with the work in speech visualization.

He has been with Oki Electric Industry Co., Ltd., Tokyo, Japan, since 1996, where he is currently engaged in the development of automatic teller machines.

Masahiro Nakatake received the B.E. and M.E. degrees in electrical engineering and computer science, Kumamoto University, Kumamoto, Japan, in 1995 and 1997, respectively. His work was in speech visualization.

He has been with Fujitsu Oita Software Laboratory, Oita, Japan, since 1997. He is currently engaged in development of point-of-sale (POS) systems.