Acoust. Sci. & Tech. **23**, 3 (2002)

PAPER

Human language identification with reduced segmental information

Masahiko Komatsu[1,2,*], Kazuya Mori[1], Takayuki Arai[1], Makiko Aoyagi[3] and Yuji Murahara[1]

[1]*Department of Electrical and Electronics Engineering, Sophia University, 7–1, Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*
[2]*Department of Linguistics, University of Alberta, 4–32 Assiniboia Hall, Edmonton, Alberta, T6G 2E7, Canada*
[3]*Center for the Teaching of Foreign Languages, Sophia University, 7–1, Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*

(*Received 4 July 2001, Accepted for publication 19 January 2002*)

Abstract: We conducted human language identification experiments using signals with reduced segmental information with Japanese and bilingual subjects. American English and Japanese excerpts from the OGI Multi-Language Telephone Speech Corpus were processed by spectral-envelope removal (SER), vowel extraction from SER (VES) and temporal-envelope modulation (TEM). The processed excerpts of speech were provided as stimuli for perceptual experiments. We calculated D indices from the subjects' responses, ranging from −2 to +2 where positive/negative values indicate correct/incorrect responses, respectively. With the SER signal, where the spectral-envelope is eliminated, humans could still identify the languages fairly successfully. The overall D index of Japanese subjects for this signal was +1.17. With the VES signal, which retains only vowel sections of the SER signal, the D index was lower (+0.35). With the TEM signal, composed of white-noise-driven intensity envelopes from several frequency bands, the D index rose from +0.29 to +1.69 corresponding to the increasing number of bands from 1 to 4. Results varied depending on the stimulus language. Japanese and bilingual subjects scored differently from each other. These results indicate that humans can identify languages using signals with drastically reduced segmental information. The results also suggest variation due to the phonetic typologies of languages and subjects' knowledge.

Keywords: Language identification, Human perception, Segmentals, Suprasegmentals, Prosody, OGI Multi-Language Telephone Speech Corpus

PACS number: 43.71.Hw

1. INTRODUCTION

Language identification (LID) with suprasemgental cues is an interesting research topic for both engineers and linguists. Exploration of humans' capability of LID is not only linguistically interesting but will contribute to the development of robust automatic LID systems.

Research has shown that it is essential to use suprasegmental cues when applying automatic LID techniques to noisy environments in which segmental information is damaged. Although much of the research on automatic LID has focused on segmental information [1–5], efforts to incorporate suprasegmental information into the system have also been made [6–8]. In recent research, the combination of segmental and suprasegmental information has achieved good results [9], underscoring the importance of suprasegmental cues.

Humans have a great capacity for LID [10], and research suggests that suprasegmental information contributes to human discriminability [11,12]. Many perceptual experiments have shown that humans can discriminate languages and dialects based on suprasegmental cues to some extent [6,13–20]. These perceptual experiments have used various stimuli, such as low-pass filtered speech [13–15], laryngograph output [14,16,17], triangular and sinusoidal pulse trains simulating the fundamental frequency, amplitude and voice timing of speech [18,19], LPC-resynthesized speech holding the coefficients constant [6] and resynthesized speech preserving or degrading broad phonotactics, syllabic rhythm or intonation [20]. However,

*e-mail: koma2@splab.ee.sophia.ac.jp

with the exception of one [20], these studies do not reveal which suprasegmental features play a role in LID, and none of them investigates how such features work together with segmental information.

To address these residual questions, it is necessary to use an experimental paradigm that compares the perceptual performance of stimuli containing different information. Ramus and Mehler [20] did so by composing several stimuli from prepared units, but their method had to assume the existence of phonological units. In the present paper, we propose a method to create stimuli by reducing some information from the original speech rather than by building up stimuli from parts. Our method, therefore, dispenses with the assumption in their paradigm (see 4.6 and 4.7 for detailed discussion).

We conducted a series of perceptual LID experiments with several types of stimulus signals containing different amounts of segmental and suprasegmental information. We created the signals from American English and Japanese speech by three different methods: spectral-envelope removal (SER), vowel extraction from SER (VES) and temporal-envelope modulation (TEM). All of these signals retain some suprasegmental information and greatly reduced segmental information. These signals were used as stimuli in the perceptual experiments. We compare results from these acoustically distinct stimuli and discuss what insights they give us into which suprasegmental features play a role in LID, and how such features work together with segmental information.

Also interesting is the notion that crosslinguistically there appear to be different cues for LID. For example, languages with typologically different prosodic characteristics may provide different acoustic cues. The variation of subjects' linguistic knowledge may also affect which cues listeners use for LID. In our experiments, Japanese monolingual subjects generally showed better identification scores for Japanese stimuli than for English stimuli. To investigate whether their better performance was due to inherent differences in the languages themselves or to the subjects' knowledge of the languages, we conducted the same set of experiments with Japanese-English bilinguals.

In the experiments, we used excerpts from the OGI Multi-Language Telephone Speech Corpus (OGI_TS) [21], which is widely used for automatic LID research and has also been used for a human LID experiment [10]. Using the publicly available OGI_TS enables us to compare results with future research into automatic and human LID.

2. SIGNAL PROCESSING

2.1. Spectral-Envelope Removal (SER)

We made a signal that contains intensity and pitch by SER. In this process, the original speech signal was whitened by removing the spectral envelope using an LPC-based inverse filter. The signal was subsequently low-pass filtered.

The use of an inverse LPC filter is based on the concept of the AR model. Regarding the mechanism of speech generation as an AR model, LPC coefficients represent the parameters of the spectral envelope of the speech signal. Therefore, inverse filtering by LPC removes the spectral information of the speech and produces the output with its spectrum flattened. This output is the driving signal of the AR model and corresponds to the glottal source of speech.

Figure 1 shows a block diagram of SER. The original signal was processed by 16th-order LPC. The sampling rate was 8 kHz, and the frame was 256 points (32 ms) long and 75% overlapped, truncated by the Hamming window. The results of the LPC analysis represent the impulse response of the FIR filter, which acts as an inverse filter of the AR model. The output of the filter, the residual signal, has a flattened spectrum similar to pseudo-periodic pulses for vowels and white noise for consonants. The gain factor of the residual signal for each frame was adjusted so as to make its energy equal to that of the original signal. The residual signal was further directed into a low-pass filter of 1-kHz cutoff to ensure the spectral removal. The amplitude of the outputs was normalized among the signals using their peak values. The resultant signals were provided for the SER experiment.

2.2. Vowel Extraction from SER (VES)

We also made a VES signal to remove possible consonantal effects from the SER signal. We extracted only the vowel sections from SER as shown in Fig. 2.

We identified the vowel sections in the signal by using the phonetic labels accompanying the corpus [21,22]. In

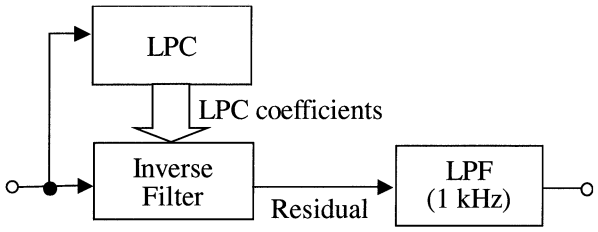


Fig. 1 Block diagram of Spectral-Envelope Removal (SER).

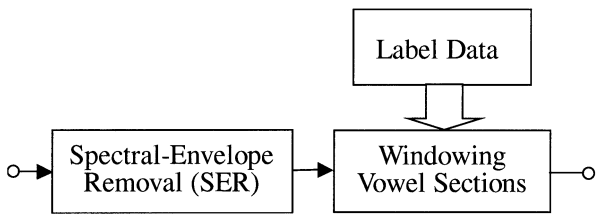


Fig. 2 Block diagram of Vowel Extraction from SER (VES).

this process, English diphthongs were treated as one vowel while Japanese hiatuses were not. Aspirated, glottalized or devoiced vowels were also included. The identified vowel sections were extracted from the SER signal by a window such that the first and the last 128 points (16 ms respectively) were the same as the first and the second halves of the Hanning window respectively and that the center portion was flat. We used this window both to avoid the clicking noise typical at the boundaries of excised speech and to reduce the possible transitional effects of consonants. 5% of the vowels were shorter than 32 ms, and they were extracted by the Hanning window instead of the above window. The consonant sections were suppressed to silence. The resultant signals were provided for the VES experiment.

2.3. Temporal-Envelope Modulation (TEM)

In TEM, we made a white-noise driven signal that retains the intensity information of several frequency bands of the original speech signal but does not include its pitch information. In this process, the temporal envelope of intensity was extracted in each of several broad frequency bands, and these envelopes were used to modulate noises of the same bandwidths. The number of bands varied from 1 to 4 as depicted in Fig. 3 (TEM 1, 2, 3 and 4), following Shannon *et al.* [23]

As an illustration, Fig. 4 shows TEM 4. The speech signal was divided into 4 signals by band-pass filters designed by the Kaiser window (transition region width: 100 Hz; tolerance: 0.001). The outputs of the band-pass filters were converted to Hilbert envelopes, which were further low-pass filtered with the cutoff at 50 Hz. These signals represent the temporal envelopes of the respective frequency bands. They were used to modulate the white noise limited by the same band-pass filters used for the speech signal, and the modulated signals were summed up. The amplitude of the signals was then normalized using their peak values. The resultant signals were provided for the TEM experiment.

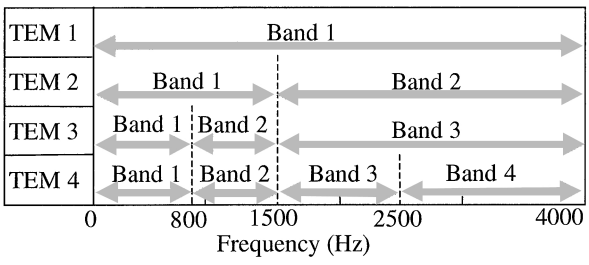


Fig. 3 Frequency division of Temporal-Envelope Modulation (TEM).

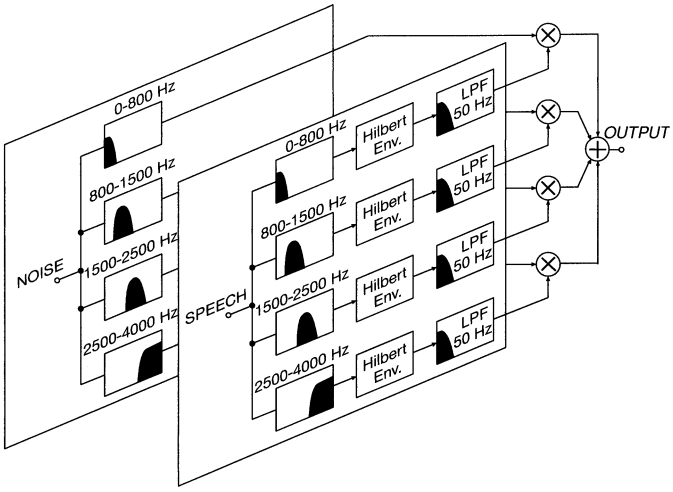


Fig. 4 Block diagram of Temporal-Envelope Modulation 4 (TEM 4). The same method was used for TEM 1, 2 and 3 except for the number of bands and their frequency division.

3. PERCEPTUAL EXPERIMENTS

3.1. Extraction of Original Utterances

We used Japanese and American English utterances from the OGI Multi-Language Telephone Speech Corpus (OGI_TS) [21]. OGI_TS is a collection of telephone speech, which includes one minute of spontaneous speech from each speaker in the corpus. We extracted two 10-second chunks of spontaneous speech from each speaker avoiding any parts with excessive hesitation, pauses, proper nouns, words of foreign origin or foreign pronunciation. 20 chunks from both males and females in both English and Japanese were extracted for a total of 80 chunks (20 chunks × 2 genders × 2 languages). Because we extracted two chunks from each speaker's speech, these 80 chunks include 40 speakers in total (10 speakers × 2 genders × 2 languages). These utterances were extracted as the input for processing by SER, VES and TEM.

3.2. Experimental Stimuli

We divided the 80 original utterances (those described in 3.1) into four data sets for the SER and TEM experiments. For VES, we chose 20 utterances out of the 80 and created only one data set. We had a smaller data set for VES for two reasons: first, because after having performed SER and TEM we did not anticipate variation among data sets; and second, because the label data had to be scrutinized before VES processing and time was at a premium. Each data set was composed of 20 stimuli, containing five males and five females in both English and Japanese. A set never included more than one utterance from the same speaker.

To prepare the stimuli for the SER experiments, the 80 original utterances (four data sets) were processed by SER

to make 80 stimuli. Each subject was provided with 20 stimuli (one data set). For each subject, a different data set was selected and the presentation order of stimuli was randomized. (For example, the first and fifth subjects took the same data set because we had only four data sets, but the stimuli were presented in different orders.)

For the VES experiments, 20 original utterances (one data set) were processed by VES to make 20 stimuli. All subjects were provided with this data set with stimuli randomized for each subject.

For the TEM experiments, the 80 original utterances (four data sets) were processed by TEM 1, 2, 3 and 4 to make 320 stimuli (80 original utterances × 4 types of TEM). Each subject was presented with 80 stimuli, a combination of four data sets respectively from TEM 1, 2, 3 and 4. To control learning effects, the combination of data sets was prepared so that each of their 80 stimuli came from a different original utterance. 24 combinations were possible under this condition. For each subject, a different combination of data sets was selected and the presentation order of stimuli was randomized. (For example, the 25th subject listened to the same 80 stimuli as one of previous subjects, but in a different order.)

See Fig. 5 for a brief description of the presentation of stimuli.

3.3. Subjects

There were 32 native speakers of Japanese (16 males and 16 females) selected independently for each of the SER, VES and TEM experiments, 96 in total (16× 2 genders × 3 methods). We also had 10 Japanese-English bilingual subjects (5 males and 5 females) for each experiment, 30 in total (5 × 2 genders × 3 methods). Out of the 30 bilingual subjects, 25 were Japanese-dominant, four were English-dominant, and one subject was equally competent in Japanese and English. We regarded as bilinguals those Japanese speakers who declared they had native-like fluency of English or who lived in English speaking regions for at least five years. Each subject voluntarily participated in the experiments (age: 18–29, average 21).

Spectral-Envelope Removal (SER)

Set 1	Set 2	Set 3	Set 4

Vowel Extraction from SER (VES)

Set 1

Temporal-Envelope Modulation (TEM)

TEM 1	Set 1	Set 2	Set 3	Set 4
TEM 2	Set 1	Set 2	Set 3	Set 4
TEM 3	Set 1	Set 2	Set 3	Set 4
TEM 4	Set 1	Set 2	Set 3	Set 4

Fig. 5 Assignment of data sets to each subject. Each subject was given either SER, VES or TEM data sets. A SER subject was given one data set (20 stimuli) out of four data sets. All VES subjects were given the same data set (20 stimuli). A TEM subject was given four data sets each from TEM 1, 2, 3 and 4 (80 stimuli in total). Shaded areas, in SER, VES and TEM respectively, indicate the data set(s) given to one subject as an example.

3.4. Procedure

The experiments were conducted using a PC in a soundproof chamber. The subject used a headset to listen to the stimuli, followed instructions on the PC display, and input the responses with a mouse. After the subject clicked the ''Play'' button on the display, a stimulus was provided binaurally through the headset. Each stimulus was presented a single time. When the headset stimulus finished playing, four buttons appeared: ''English'', ''Probably English'', ''Probably Japanese'' and ''Japanese'', from which the subject was instructed to select the most appropriate button on a forced-choice condition. No feedback was provided to the subject. After the subject made a selection, the ''Play'' button appeared for the next stimulus. A session contained either 20 SER stimuli, 20 VES stimuli or 80 TEM stimuli. The session proceeded at the subject's pace. On average, the SER experiment took approximately 10 min; the VES, 10 min; and the TEM, 30 min.

Prior to each experiment, the subject was given a practice session with four stimuli, different from those used for the actual experiment, to become familiar with the procedure. No feedback was provided for the practice session.

3.5. Experimental Results

We calculated an index of discriminability (D index) [17] averaged for each stimulus type. The D index was calculated in such a way that ''English'' and ''Japanese'' were scored as ±2 while ''Probably English'' and ''Probably Japanese'' were ±1. Positive values indicate correct responses; and negative, incorrect ones. The averaged D index ranges from −2 to +2, where 0 indicates random responses.

Figures 6 and 7 show the D indices of either subject group for SER, VES and TEM 1, 2, 3 and 4 with utterance categories, English male, English female, Japanese male and Japanese female (''Em'', ''Ef'', ''Jm'' and ''Jf'', respectively). ''All'' indicates the overall D index, which is the average of these four categories.

Japanese subjects showed an overall D index of 1.17 for SER signals, those retaining the information of the temporal envelopes of intensity and pitch. The index went

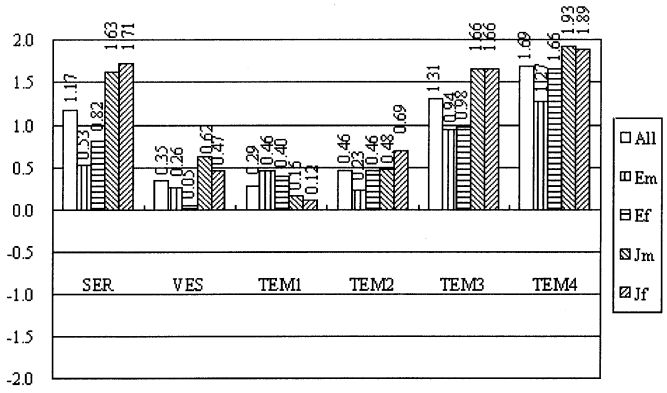


Fig. 6 D indices of Japanese subjects.

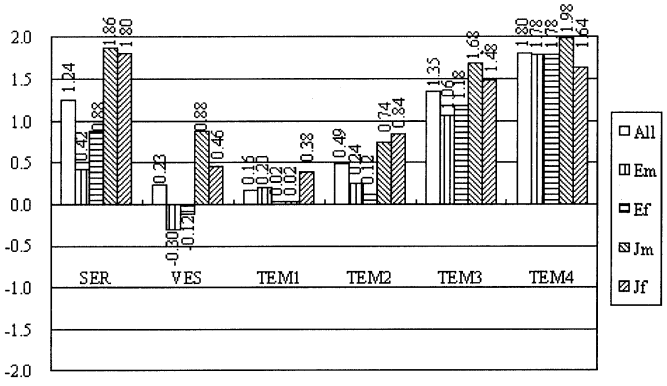


Fig. 7 D indices of bilingual subjects.

down to 0.35 for VES, which has less information. For TEM, the overall D index rose from 0.29 to 1.69 as the number of bands increased from 1 to 4. Bilingual subjects showed a similar tendency: 1.24 for SER, 0.23 for VES, 0.16 to 1.80 for TEM.

Tables 1 and 2 show both the D index and the correct identification rate for each stimulus type and language, providing rough correspondence of D indices to identification rates. The identification rates, which may be more familiar to most readers, are calculated based on the subjects' combined judgments of ''English'' and ''Probably English'' counted as the judgment for the English language as well as ''Japanese'' and ''Probably Japanese'' counted as Japanese.

Table 1 D indices and correct identification rates of Japanese subjects.

	D index		Identification rate	
Stimulus	Eng	Jpn	Eng	Jpn
SER	0.67	1.67	73.1%	97.2%
VES	0.16	0.54	55.0%	66.9%
TEM 1	0.43	0.14	68.4%	56.6%
TEM 2	0.34	0.58	60.6%	68.4%
TEM 3	0.96	1.66	79.1%	93.4%
TEM 4	1.47	1.91	89.4%	98.1%

Table 2 D indices and correct identification rates of bilingual subjects.

	D index		Identification rate	
Stimulus	Eng	Jpn	Eng	Jpn
SER	0.65	1.83	70.0%	100.0%
VES	−0.21	0.67	44.0%	66.0%
TEM 1	0.11	0.20	57.0%	60.0%
TEM 2	0.18	0.79	56.0%	78.0%
TEM 3	1.12	1.58	79.0%	91.0%
TEM 4	1.78	1.81	95.0%	96.0%

4. DISCUSSION

4.1. SER (Japanese Subjects)

Our results show a distinctive difference in identification between languages but not between genders. The overall D index for the Japanese stimuli (1.67) is higher than that for the English stimuli (0.67). Note that the English result is still high enough to be considered successful LID.

Post-experimental questionnaires indicated the subjects used differences in intonation between the languages as a cue to LID. Some peculiar intonation contours at phrase junctures seem to provide clues, suggesting one language over the other. The spontaneous speech in the OGI_TS we used is monologue, in which English speech has a lot of rising intonation at phrase junctures while Japanese often presents a different type of lengthened, rise-fall intonation in the phrase endings. These intonational characteristics are certainly detectable in the SER signal, which acoustically contains intensity and F_0 contour information.

The questionnaires indicated that subjects could often detect Japanese words in the utterances. Although SER does not retain spectral envelope information, incomplete phonotactic information is still present. The vowel/consonant distinction and identification of the manner of articulation is realized by the existence of harmonics or white noise and the temporal change of intensity. As it is not clear from our results whether or not the subjects actually recognized words, we are inclined to believe that at times phonotactic information enabled the subject to spot words, and at other times subjects only imagined they were hearing words.

4.2. VES (Japanese Subjects)

With VES, there was also a distinctive difference in D indices between the languages (Japanese: 0.54; English: 0.16) but not between the genders. The overall D index for VES (0.35) was lower than SER (1.17) and may be regarded as an unsuccessful result. The index for only Japanese stimuli may be regarded as barely successful

(0.54).

Note that Ef (0.05) for VES is quite low and lower than Em (0.26). However, we would emphasize that English scores for VES are generally low and not admit that the difference between Ef and Em for VES is distinctive. It is of note also that not only are the differences between genders small but gender has no consistent effect in the experiments as a whole. For example, the male stimuli are slightly better than the female for VES, but vice versa for SER.

The VES signal keeps intonational information as does the SER signal. The post-experimental questionnaires indicated that the subjects used intonation as a cue to LID with VES, just as they did for SER.

The VES signal is different from the SER signal in that it has no consonant sections. We suspected that the removal of consonant information deprived subjects of word-spotting strategies, and this was confirmed by the questionnaires. We believe this difference accounts for the lower VES index.

4.3. TEM (Japanese Subjects)

With TEM, just as with SER and VES, there was a distinctive difference in D indices between the languages but not between the genders. The overall D indices for the Japanese stimuli (TEM 1–4: 0.14, 0.58, 1.66 and 1.91) were better than their English counterparts (TEM 1–4: 0.43, 0.34, 0.96 and 1.47), but the difference was not as marked as in SER. For the Japanese stimuli, the indices rise as the number of bands increases. For the English stimuli, there is not a noticeable difference between the indices of TEM 1 and 2 though an increase is certainly observed from TEM 2 to 4.

In TEM 1, though we do not regard these indices as successful results of LID, the English indices are higher than the Japanese ones. As the TEM 1 signal carries only the information on the temporal change of intensity, we compared the modulation spectra of our original utterances in both languages. Figure 8 shows the impulse responses (in zero phase) obtained from the modulation spectra. Though there is no distinctive difference in general as was also pointed out by Arai and Greenberg [24], English has a larger drop around 250 ms than Japanese. This difference may have caused the higher scores for English.

The ascending tendency of the scores along with the increasing number of bands conforms to the results by Shannon *et al.* [23], who conducted speech recognition experiments with almost the same signals as TEM. Their results indicate that segments are more correctly identified as the number of bands increases. Our results of LID are especially similar to their result of a sentence recognition task in the respect that there is a jump from the 2-band to the 3-band conditions (1- to 4-band conditions: 3%, 27%, 84% and 96%; these values are read from their graph). For an analogy in Japanese speech, we may refer to Obata and Riquimaroux [25], who discuss the proper frequency division for the perception of Japanese vowels. (They argue that 600 Hz and 1,500 Hz are the best dividing points in the 3-band condition.) In our post-experimental questionnaires, the subjects responded that they used word-spotting strategies far more often than intonational cues in the TEM experiments. The foregoing findings altogether suggest that segmental cues are the most convincing explanation for the increase in the scores from TEM 1 to 4.

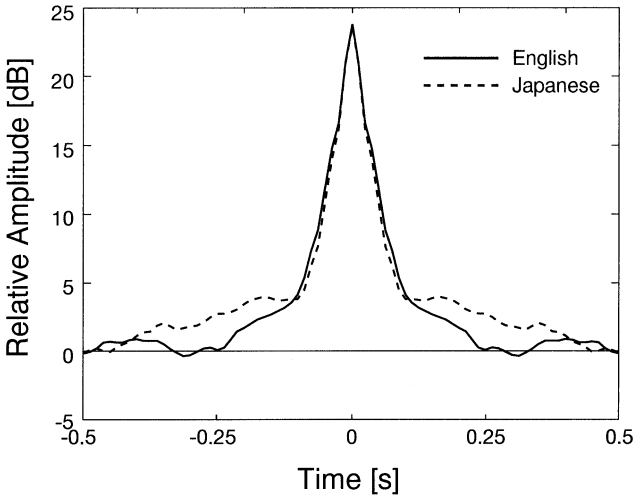


Fig. 8 Impulse responses of modulation (in zero phase).

4.4. Bilingual Subjects

Bilingual subjects showed results similar to the Japanese subjects. Of note is that bilingual subjects generally registered higher D indices for Japanese rather than English stimuli, just as the Japanese subjects did. This finding suggests that the different D indices between English and Japanese may be attributable to the phonetic differences inherent in the two languages rather than the subjects' linguistic knowledge of these languages. However, we will not be certain of this conclusion until we have data from English monolinguals.

We regard that the difference between genders is not distinctive generally in the results from the bilingual subjects. For example, the scores for Ef and Jm of TEM 1 in Fig. 7 are quite low, but the scores for the opposite gender counterparts, Em and Jf, are also low. Hence, we surmise that the overall English and Japanese scores for TEM 1 are low, and do not consider that the scores for Ef and Jm are especially lower than Em and Jf respectively.

It is also interesting that there are differences between the subject groups, as seen in Fig. 9. ''Eng'' and ''Jpn'' in Fig. 9 indicate the language of the stimuli, while ''Mono'' and ''Bi'' indicate the subject groups. In the graph, VES and TEM 1 are the stimuli that have no, or little if any,

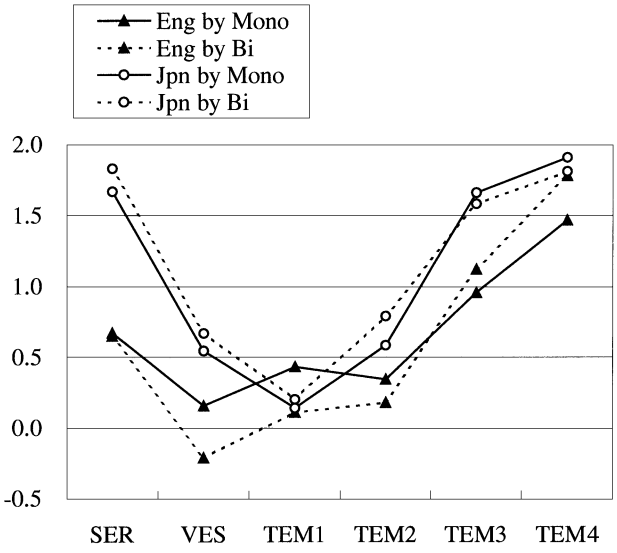


Fig. 9 D indices of Japanese and bilingual subjects.

segmental information, and the amount of additional information increases when it goes to either side of the graph. The difference between the subject groups is generally larger for the English stimuli than for the Japanese stimuli. This makes sense because the variation of subjects' linguistic knowledge of English is greater than that of Japanese. The steeper increase in the scores of TEM 2–4 for English stimuli by the bilingual subjects than by the Japanese subjects suggests that the bilingual subjects were able to use the segmental information more efficiently than the Japanese subjects. The scores of VES and TEM 1–2 for English stimuli, on the other hand, suggest that the Japanese subjects used suprasegmental information more efficiently than the bilingual subjects. Therefore, we believe that listeners with different linguistic knowledge use different acoustic cues for LID.

4.5. Reduction in Segmental Information and the Role of Partial Phonotactics

There are investigations on the recognition of segments in the LPC residual signals which maintain the intensity contour of the original speech signal. In Arai *et al.*'s experiments [26], the average correct identification rate of Japanese consonant-vowel (CV) syllable in the LPC residual was 18.5% (chance level: 3.2%) or 30.6% (chance level: 4.8%) depending on the choice of syllables included in the stimulus set. In Komatsu *et al.* [27], the identification rate was 20.0% (chance level: 5.9%) for consonants, 46.9% for their manner features, 30.8% for place features, 81.4% for voice features and 66.4% for major categories (discrimination of obstruent/nasal/liquid/glide). Place features, which need fine spectral structure for recognition, showed an especially lower rate than the other features. Also note that these features are not independent from each other: e.g., if the manner of a consonant is ''liquid,'' its place and voice must be ''alveolar'' and ''voiced.'' Komatsu *et al.* [28] showed the identification rates of consonants in consonant clusters differ from those in CV syllables, as they are affected by language-dependent phonotactic constraints. The experimental results [27,28] also suggested variation due to listeners' first languages. These analyses support the argument that the LPC residual effectively suppresses information used in phoneme identification, but not information needed to distinguish broad categories of phonemes, i.e., ''sonority.'' Sonority provides important cues to the syllable structures of a language (see [29] for detailed discussion). Because SER is the low-pass filtered LPC residual, we assume such properties of the LPC residual are inherited by the SER signal.

The fact that VES had a lower index than SER is consistent with the argument by Ramus and Mehler [20] that syllabic rhythm is an essential cue to LID. Since there is enough information in the SER signal to reconstruct the basic phonotactic structure, the syllable structure becomes evident; and according to Ramus, Nespor and Mehler [30], syllabic structure plays a role in creating syllabic rhythm. Meanwhile in VES, because consonants are all replaced with silence, there is no segmental or phonotactic clue that contributes to syllable structure recognition. The temporal pattern of silence-vowel alternation in VES was not sufficient to create the sense of syllabic rhythm, and LID was degraded.

Shannon *et al.* [23] reports comparatively high phoneme recognition rates in the experiments with almost the same signals as TEM. Their recognition scores in the 1- to 4-band conditions were approximately 48%, 69%, 78% and 89% (chance level: 6.3%) for consonants; and 35%, 62%, 88% and 95% (chance level: 13.5%) for vowels. They also report analysis by consonantal feature: 67%, 90%, 93% and 94% for voice; 77%, 95%, 95% and 97% for manner; and 16%, 31%, 45% and 64% for place. (These values are read from their graphs.) Even in the 1-band condition, the consonant identification rated as high as 48%. However, we cannot assume based on this score that our subjects could identify phonemes as well in the TEM 1 signal. Experimental results vary largely depending on experimental conditions. In Shannon *et al.*'s 1- and 2-band conditions, the recognition task for consonants and vowels, where the subjects listened to /vCv/ and /cVc/, showed much higher scores than the sentence recognition task (mentioned in 4.3). This suggests that identification of individual phonemes is not as accurate in connected speech. Likewise, accurate recognition of single phonemes does not imply accurate recognition of phoneme sequences (phonotactics), most likely due to coarticulation effects. So far, we do not know how phoneme clusters can be

identified in spectrally reduced signals such as TEM. (As to the LPC residual, it has been shown that the recognition of consonants in consonant clusters differs from that in CV syllables [28].) We cannot say TEM 1 has more segmental information than SER by merely comparing Shannon *et al.*'s results with the results for the LPC residual [26–28], drawn from different experimental settings. Considering the performance is rather unsuccessful both in Shannon *et al.*'s sentence recognition in the 1-band condition and in our TEM 1, we infer that there is not much segmental information in TEM 1, if there is any at all.

We consider phonotactics is especially an important contributor to LID improvements from TEM 1 to 4. Typologically frequent occurrence of consonant clusters makes English eminently different from Japanese. Also, discrimination of single phonemes in a language potentially improves LID only to the extent that the inventories of the languages are different. Mere recognition of /t/ in a stimulus does not help distinguish between English and Japanese because both have a /t/ sound. Discrimination of subtle acoustic cues (e.g., dental place in Japanese /t/ vs. alveolar place in English /t/) may well improve LID if listeners had enough information to discriminate these differences. However, considering that even TEM 4 is spectrally quite distorted, we do not imagine that subjects are able to detect such subtle acoustic differences between phonemes. Rather, we assume partial recognition enables listeners to reconstruct partial phonotactic information, which then enables LID.

4.6. Segmental/Suprasegmental Information Contained in the Signal

We summarized the information contained in the stimuli and the results of LID for each of our experiments in Table 3. In the table, we define ''segmental'' as the information represented by the frequency spectrum and ''suprasegmental'' as the information represented by temporal contours. (We will discuss the problem this division implies later.)

The descriptions of segmental information in Table 3 are what we speculate from the signal processing methods we used. The descriptions are not necessarily a reflection of the recognition scores in the literature ([23,26–28] mentioned in 4.5) because of the different experimental conditions. We parenthesized the segmental information in TEM 1 because it is not induced from spectral information even if there is some segmental information in TEM 1. As already discussed in 4.5, we consider ''phonotactics'' (even if incomplete) as more important than ''acoustics'' (both broad and fine phonetic) in our experiments.

Intensity information is available in all of these signals, and pitch information is available in only SER and VES. Note that in VES such information is available only for vowel sections because consonants are all suppressed to silence.

Although we divide information into two large categories (segmental vs. suprasegmental) in Table 3, for simplicity, this dichotomy by acoustic properties (spectral vs. temporal) is approximate and implies theoretical problems. First, not only spectral but temporal property contributes to the perception of segments. This is clear when we see the recognition scores of phonemes higher than chance levels in the 1-band condition in Shannon *et al.*'s experiments [23] (see 4.5) where only the intensity contour is available. Second, recent research [20,30] shows the importance of factors such as syllable structures in the perception of rhythm, which should be regarded as suprasegmental. Because syllable structure is linguistically determined by the sequence of broad categories of segments, some segmental information is necessary to create a sense of rhythm (see [29] for detailed discussion). We do not claim to have completely separated suprasegmentals from segmentals in our experiments. Rather, we assume complete separation is impossible, and in the present research we gradually ''rub off'' segmental properties from speech.

4.7. Comparison with Other Studies

Our experimental design may look similar to that of Ramus and Mehler [20], but their approach is different. They conducted perceptual experiments on English and

Table 3 Information contained in the signal and the results of human LID.

	Segmental (Spectral)		Suprasegmental (Temporal)		Results of LID
	Acoustics	Phonotactics	Intensity	Pitch	
SER	little	only broad	available	available	successful
VES	none		available	available	unsuccessful
TEM 1	(little)		available	none	unsuccessful
TEM 2	⋮				⋮
TEM 3	increasing but much reduced				getting better
TEM 4	⋮				⋮
					successful

Japanese. They phonemically segmented the original English and Japanese speech and then replaced the segments with French phonemes to exclude segmental cues to LID. In replacing, they substituted all fricatives with /s/, all stops with /t/ and so on for one type of stimuli; all consonants with /s/ and so on for another type; and the like. Thus, they created four types of stimulus signals including or excluding broad phonotactics, syllabic rhythm and intonation. In their experimental process, they assume some phonological unit, which they segment speech into, and operate on this unit in order to separate suprasegmentals from segmentals. In contrast, our method does not assume any segmental unit, and it directly operates on the continuous change of acoustics in speech. There may well be cases that phonological units should not be assumed for some research purposes. Thus, while they ''built up'' stimuli from prepared phonological units, which enabled them to control segmental and suprasegmental features separately, we chose to ''rub off'' some features from the original speech. Ramus and Mehler are not pursuing the same end. We may call their approach more ''phonological'' and ours more ''phonetic.''

Another study that is comparable to ours is Ohala and Guilbert's [18], which demonstrates the importance of suprasegmental information in LID. They used the signal that represents the fundamental frequency and amplitude of voiced sections in speech, which is close to the SER and VES signals. They are different in that their signal does not have a noise source: voiceless sections were suppressed to silence, but voiced consonants were represented as pulse trains which keep the fundamental frequency and amplitude of the original speech. Their signal may be regarded as something between SER, which has noise source for consonants, and VES, which has no consonant sections.

Miura *et al.* [31] and Ohyama and Miura [32] argue for the dominance of prosodic parameters over spectral parameters in judgments of the naturalness of English and Japanese speech. They tested with speech PARCOR-resynthesized with prosodic and spectral parameters from native and non-native speakers combined. Especially interesting is that duration and fundamental frequency are more important than intensity. This may provide some insight into the low scores of TEM 1 in our experiments. Also interesting is that the experiments with Chinese showed different results, suggesting that the important cues differ depending on the prosodic system of the language.

In automatic LID research, Nakagawa *et al.* [33] shows the effectiveness of HMM state sequences. Their findings are comparable to ours in the discussion of broad phonotactics. Considering that results for SER were much better than for VES, it is clear that broad phonotactic information is useful for human LID, as well as for automatic LID, as the Nakagawa study indicated. It is also plausible that the results improved from TEM 1 to 4 as more phonotactic information was available. Thus, both Nakagawa *et al.* and the present study show that phonotactic information is useful for LID, even if in partial form.

5. CONCLUSIONS

The temporal attributes of intensity and pitch do not alone enable LID. However, LID is possible if other information is made available, even with a greatly degenerated signal. The TEM 1 signal contains only the intensity envelope, and VES, the intensity and pitch of vowel sections. Neither TEM 1 nor VES provided the subjects with sufficient information to identify the languages. In SER phonotactic information (albeit incomplete) combined with intensity and pitch information enabled better LID. In TEM 2–4 the D index rose as the number of bands increased, and here segmental information was an important contributing factor.

Our results showed the importance of suprasegmental information when combined with a small amount of segmental information such as broad phonotactics. From our results we cannot conclude that LID is possible solely based upon the suprasegmental information. Instead, we argue that the suprasegmental information, specifically intensity and pitch, can be used under conditions where the segmental information, the acoustics of segments and phonotactics, is severely reduced. This is confirmed when we see the high D indices of SER (Japanese subjects: 1.17; bilingual subjects: 1.24), where the segments are severely degenerated while the suprasegmental attributes of intensity and pitch are still present. The nearly perfect scores for TEM 4 (Japanese subjects: 1.69; bilingual subjects: 1.80) also supports our claim, because the segmental information is greatly reduced with TEM, yet LID was possible. We especially recognize the importance of phonotactic information which is partially kept when the signal is spectrally reduced.

Bilingual subjects showed the same general tendency as monolingual subjects. The importance of suprasegmental information combined with reduced segmental information was confirmed for both monolinguals and bilinguals.

Our experiments also suggest that different cues to LID are available depending on the language. English and Japanese have different accent, rhythm and syllable structures. We suspect that such different typologies result in an unequal availability of prosodic cues for the two languages, which results in different D indices.

The results also suggest that listeners' linguistic knowledge affects the cues they use. Not all potential cues are perceived by all listeners. Rather, subjects seem to have limited access to the cues according to their linguistic knowledge. If subjects are given extensive training with

feedback, they may grow more sensitive to the signals and be able to utilize more cues, resulting in higher identification scores for both languages. Thus the signals used in these experiments may embody more clues to LID than revealed at this time.

ACKNOWLEDGMENTS

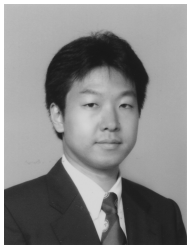
This article is the revised version of the papers presented at Eurospeech 99 and 2001 [34,35]. We are grateful for the Foundation ''Hattori-Hokokai'' for their grant in 1999. We also thank Noriaki Toba and Tomoyuki Harada for their collaborative work at an early stage of the research, Terri Lander for reading the draft for improvement and two anonymous reviewers for their comments.

REFERENCES

[1] Y. K. Muthusamy, E. Barnard and R. A. Cole, ''Reviewing automatic language identification'', *IEEE Signal Process. Mag.*, **11**(4), 33 (1994).
[2] Y. K. Muthusamy, K. Berkling, T. Arai, R. A. Cole and E. Barnard, ''A comparison of approaches to automatic language identification using telephone speech'', *Proc. Eurospeech*, **93**, 1307 (1993).
[3] T. Arai, ''Automatic language identification using sequential information of phonemes'', *IEICE Trans.*, **E78-D**, 705 (1995).
[4] Y. Yan, E. Barnard and R. A. Cole, ''Development of an approach to automatic language identification based on phone recognition'', *Comput. Speech Lang.*, **10**, 37 (1996).
[5] M. A. Zissman, ''Comparison of four approaches to automatic language identification of telephone speech'', *IEEE Trans. Speech Audio Process.*, **4**, 31 (1996).
[6] J. T. Foil, ''Language identification using noisy speech'', *Proc. ICASSP*, **86**, 861 (1986).
[7] M. Savic, E. Acosta and S. K. Gupta, ''An automatic language identification system'', *Proc. ICASSP*, **91**, 817 (1991).
[8] A. E. Thymé-Gobbel and S. E. Hutchins, ''On using prosodic cues in automatic language identification'', *Proc. ICSLP*, **96**, 1768 (1996).
[9] S. Itahashi, T. Kiuchi and M. Yamamoto, ''Spoken language identification utilizing fundamental frequency and cepstra'', *Proc. Eurospeech*, **99**, 383 (1999).
[10] Y. K. Muthusamy, N. Jain and R. A. Cole, ''Perceptual benchmarks for automatic language identification'', *Proc. ICASSP*, **94**, 333 (1994).
[11] Z. S. Bond, D. Fucci, V. Stockmal and D. McColl, ''Multidimensional scaling of listener responses to complex auditory stimuli'', *Proc. ICSLP*, **98**, Vol. 2, 93 (1998).
[12] V. Stockmal, D. R. Moates and Z. S. Bond, ''Same talker, different language'', *Proc. ICSLP*, **98**, Vol. 2, 97 (1998).
[13] K. Atkinson, ''Language identification from nonsegmental cues [Abstract]'', *J. Acoust. Soc. Am.*, **44**, 378 (1968).
[14] A. Moftah and P. Roach, ''Language recognition from distorted speech: Comparison of techniques'', *J. Int. Phonet. Assoc.*, **18**, 50 (1988).
[15] R. Mugitani, A. Hayashi and S. Kiritani, ''Yoiku-kankyo ni aru hogen e no senko-hanno no hattatsu [Developmental change of 5 to 8-month-old infants' preferential listening response for the native dialect]'', *J. Phonet. Soc. Jpn.*, **4**(2), 62 (2000).
[16] J. A. Maidment, ''Voice fundamental frequency characteristics as language differentiators'', *Speech and Hearing: Work in Progress*, **2**, 74 (University College London, 1976).
[17] J. A. Maidment, ''Language recognition and prosody: Further evidence'', *Speech, Hearing and Language: Work in Progress*, **1**, 133 (University College London, 1983).
[18] J. J. Ohala and J. B. Gilbert, ''Listeners' ability to identify languages by their prosody'', in *Problèmes de Prosodie: Vol. 2, Expérimentations, Modèles et Fonctions*, P. Léon and M. Rossi, Eds. (Didier, Paris, 1979), p. 123.
[19] M. Barkat, J. Ohala and F. Pellegrino, ''Prosody as a distinctive feature for the discrimination of Arabic dialects'', *Proc. Eurospeech*, **99**, 395 (1999).
[20] F. Ramus and J. Mehler, ''Language identification with suprasegmental cues: A study based on speech resynthesis'', *J. Acoust. Soc. Am.*, **105**, 512 (1999).
[21] Y. K. Muthusamy, R. A. Cole and B. T. Oshika, ''The OGI Multi-Language Telephone Speech Corpus'', *Proc. ICSLP*, **92**, 895 (1992).
[22] T. Lander, *The CSLU Labeling Guide* (Center for Spoken Language Understanding Technical Report, No. CSLU-014-96, Oregon Graduate Institute of Science and Technology, 1996).
[23] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski and M. Ekelid, ''Speech recognition with primarily temporal cues'', *Science*, **270**, 303 (1995).
[24] T. Arai and S. Greenberg, ''The temporal properties of spoken Japanese are similar to those of English'', *Proc. Eurospeech*, **97**, 1011 (1997).
[25] Y. Obata and H. Riquimaroux, ''Role of temporal information in speech perception: Importance of amplitude envelope information'', *Proc. Spring Meet. Acoust. Soc. Jpn.*, p. 369 (1999).
[26] T. Arai, M. Pavel, H. Hermansky and C. Avendano, ''Syllable intelligibility for temporally filtered LPC cepstral trajectories'', *J. Acoust. Soc. Am.*, **105**, 2783 (1999).
[27] M. Komatsu, W. Tokuma, S. Tokuma and T. Arai, ''The effect of reduced spectral information on Japanese consonant perception: Comparison between L1 and L2 listeners'', *Proc. ICSLP*, **2000**, Vol. 3, 750 (2000).
[28] M. Komatsu, T. Shinya, M. Takasawa and T. Arai, ''LPC zansa-shingo ni yoru shiin-ninshiki to onso-hairetsu [Consonant perception in LPC residuals and phonotactics]'', *Proc. 14th Gen. Meet. Phonet. Soc. Jpn.*, p. 49 (2000).
[29] M. Komatsu, ''What constitutes acoustic evidence of prosody? The use of Linear Predictive Coding residual signal in perceptual language identification'', *LACUS Forum*, **28** (Linguistic Association of Canada and the United States, in press).
[30] F. Ramus, M. Nespor and J. Mehler, ''Correlates of linguistic rhythm in the speech signal'', *Cognition*, **73**, 265 (1999).
[31] I. Miura, G. Ohyama and H. Suzuki, ''A study of the prosody of Japanese English using synthesized speech'', *Proc. Autumn Meet. Acoust. Soc. Jpn.*, p. 239 (1989).
[32] G. Ohyama and I. Miura, ''A study on prosody of Japanese spoken by foreigners'', *Proc. Spring Meet. Acoust. Soc. Jpn.*, p. 263 (1990).
[33] S. Nakagawa, T. Seino and Y. Ueda, ''Spoken language identification by Ergodic HMMs and its state sequences'', *IEICE Trans.*, **J77-A**, 182 (1994).
[34] K. Mori, N. Toba, T. Harada, T. Arai, M. Komatsu, M. Aoyagi and Y. Murahara, ''Human language identification with reduced spectral information'', *Proc. Eurospeech*, **99**, 391 (1999).
[35] M. Komatsu, K. Mori, T. Arai and Y. Murahara, ''Human language identification with reduced segmental information: Comparison between monolinguals and bilinguals'', *Proc. Eurospeech*, **2001**, 149 (2001).

Masahiko Komatsu (M.A. in Linguistics 1987, B.A. in Foreign Studies 1985 both from Sophia University, Tokyo) is currently a Ph.D. student of the Department of Linguistics, University of Alberta, Canada, and a collaborating researcher of the Department of Electrical and Electronics Engineering, Sophia University. In 1991–1996, he was an assistant professor of the Department of English Literature, Beppu University. In 1987–1990, he worked for AI Section, Kozo Keikaku Engineering Inc., Tokyo, chiefly engaged in developing an algebraic specification language ASPELA. His current interest is in prosodic typology of languages.

Kazuya Mori received the B.E. and M.E. degrees in electrical engineering from Sophia University, Tokyo, Japan, in 1998 and 2000, respectively. He engaged in all of the experiments in this paper during his master's studies. He now works for a national network, Tokyo Broadcasting System, Inc. (TBS), Tokyo, Japan, as an acoustic technician.



Takayuki Arai received the B.E., M.E. and Ph.D. degrees in electrical engineering from Sophia Univ., Tokyo, Japan, in 1989, 1991 and 1994, respectively. In 1992–1993 and 1995–1996, he was with Oregon Graduate Institute of Science and Technology (Portland, OR, USA). In 1997–1998, he was with International Computer Science Institute and Univ. of California (Berkeley, CA, USA). He is currently Associate Professor of the Department of Electrical and Electronics Engineering, Sophia Univ. He was a short-term visiting scientist at several institutions, including M.I.T. and Max Planck Institute for Psycholinguistics. His research interests include acoustics, speech and hearing sciences, and spoken language processing.

Makiko Aoyagi received M.A. (1992) and Ph.D. candidacy (2000) in linguistics from Sophia University, Tokyo. The area of her interest is phonetics/phonology, and her current research topic is coarticulation phenomena. She is currently Assistant Professor of the Foreign Language Center, Sophia University.

Yuji Murahara received B.E. in electronics engineering in 1963 from the University of Electro-Communications, and received Ph.D. in electrical engineering in 2001 from Sophia University. From 1963 to March 2002 he was Research Assistant in the Department of Electrical and Electronics Engineering, Sophia University, and is now Lecturer there. His research interests include digital signal processing and biomedical engineering.