Human Language Identification with Reduced Segmental Information: Comparison between Monolinguals and Bilinguals

Masahiko Komatsu[1,2], *Kazuya Mori*[1,*], *Takayuki Arai*[1], *Yuji Murahara*[1]

[1] Department of Electrical and Electronics Engineering
Sophia University, Tokyo, Japan
http://www.splab.ee.sophia.ac.jp/
[2] Department of Linguistics
University of Alberta, Edmonton, Canada

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_TS Corpus were processed by spectral-envelope removal (SER), vowel extraction from SER (VES) and temporal-envelope modulation (TEM). With the SER signal, where the spectral-envelope is eliminated, humans could still identify the languages fairly successfully. With the VES signal, which retains only vowel sections of the SER signal, the identification score was low. With the TEM signal, composed of white-noise-driven intensity envelopes from several frequency bands, the identification score rose as the number of bands increased. Results varied depending on the stimulus language. Japanese and bilingual subjects demonstrated different scores from each other. These results indicate that humans can identify languages using a signal with drastically reduced segmental information. The results also suggest variation due to the phonetic attributes of languages and subjects' knowledge.

1. Introduction

Language identification (LID) with suprasemgental cues provides 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 a robust automatic LID system.

Applying the automatic LID technique to a noisy environment where segmental information is damaged requires using suprasegmental cues. Although much of the research on automatic LID has focused on segmental information [1], the combination of segmental and suprasegmental information has achieved a good result [2], revealing the importance of suprasegmental cues.

Humans have a great capacity of LID [3]. Perceptual experiments have shown that humans can discriminate languages and dialects based on suprasegmental cues to some extent [4][5][6]. It is also pointed out that broad phonotactic information, the temporal alignment of manner-of-articulation features of phonemes, significantly helps [7].

Our previous perceptual experiments also showed the importance of suprasegmental information when combined with a small amount of segmental information [8]. We conducted the experiments on Japanese monolinguals with English and Japanese speech signals that had been processed by spectral-envelope removal (SER) and temporal-envelope modulation (TEM). The temporal change of intensity and pitch was not enough for LID by itself; but if other information was added, LID was quite possible, even with a significantly degraded signal. We argued that suprasegmental information, specifically intensity and pitch, can be used under conditions where segmental information, in particular the acoustics of segments and phonotactics, is severely reduced.

In our perceptual experiments, the identification score for Japanese stimuli was generally higher than for English stimuli. It was not clear whether this was due to inherent differences in the languages themselves or to the subjects' varying knowledge of these languages.

In the present study, to investigate this further, we conducted follow-up experiments on Japanese-English bilinguals, meanwhile adding another type of stimulus: vowel extraction from SER (VES). In this paper, we discuss these effects based on the results of our series of experiments.

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.

Fig. 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 msec) 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 eliminate any spectral information that may still remain. 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 the 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 phonemic labels accompanying the corpus [9][10]. In this process, English diphthongs were treated as one vowel while Japanese hiatuses were not. The identified vowel sections were extracted from the SER signal by such a window that the first and the last 128 points (16 msec) were the same as the first and last halves of the Hanning window respectively, and the center portion was flat. We used this window both to avoid the clicking noise at the beginning and end of the section and to reduce the possible transitional effects of consonants. 5% of the vowels were shorter than 32 msec, 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.

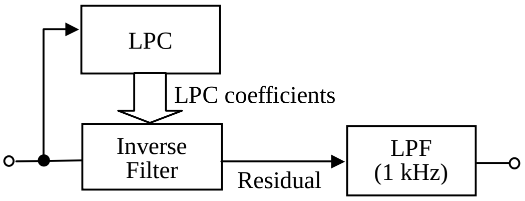

Figure 1: Block diagram of SER.

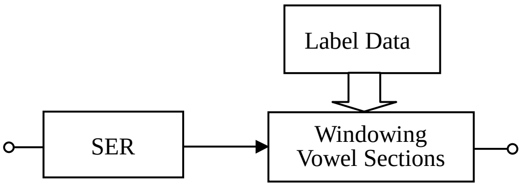

Figure 2: Block diagram of VES

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. [11].

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 modulated by the white noise limited by the same band-pass filters used for the speech signal, and summed up. The amplitude of signals thus produced was normalized using their peak values. The resultant signals were provided for the TEM experiment.

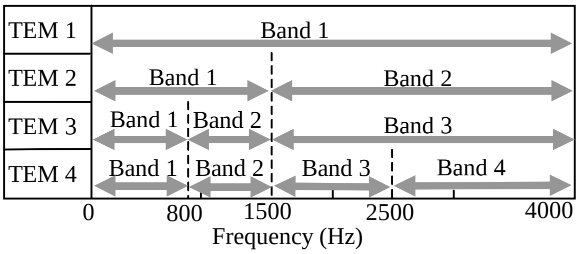

Figure 3: Frequency division of TEM

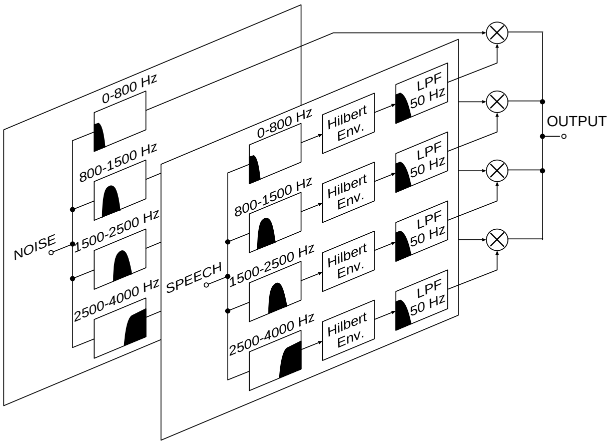

Figure 4: Block diagram of TEM 4

3. Perceptual experiments

3.1. Extraction of original utterances

We used Japanese and English utterances from the OGI Multi-Language Telephone Speech Corpus (OGI_TS) [9]. 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). These utterances were extracted as the input for processing by SER, VES and TEM.

3.2. Experimental stimuli

To prepare the stimuli for the SER experiment, the 80 original utterances (those described in 3.1) were processed by SER to make 80 stimuli. We divided the 80 stimuli into 4 data sets so that each data set was composed of 20 stimuli, containing 5 each of English male/female and Japanese male/female. For each subject a different data set was selected, and the arrangement within the data set was randomized for each subject.

For the VES experiment, 20 original utterances (5 each from English male/female and Japanese male/female) were randomly selected out of the 80 original utterances and processed by VES to make 20 stimuli. Only one data set was prepared, which was composed of these 20 stimuli. For each subject, the arrangement within the data set was randomized.

For the TEM experiment, the 80 original utterances 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 selected from the 320. The 80 stimuli were composed of 20 each from TEM 1, 2, 3 and 4, where each TEM group contained 5 each of English male/female and Japanese male/female. To control for learning effects, the data set was prepared so that each of its 80 stimuli came from a different original utterance. For each subject a different data set was selected, and the arrangement in each data set was randomized.

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). Each subject participated voluntarily in the experiments (age: 18-29, average 21).

3.4. Procedure

The experiments were conducted in a soundproof chamber, using a PC. 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 through the headset. Each stimulus was presented only a single time. When the headset stimulus finished playing, 4 buttons appeared: "English", "Probably English", "Probably Japanese" and "Japanese", from which the subject was instructed to select the most appropriate button. No feedback was provided. After the subject made a selection, the "Play" button appeared for the next stimulus. The session contained either 20 SER stimuli, 20 VES stimuli, or 80 TEM stimuli. The session proceeded at the subject's own pace. On average, the SER or VES experiment took approximately 10 minutes, and the TEM took approximately 30 minutes.

Prior to the experiment discussed above, the subject was given a practice session with 4 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) [4] 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.

Figs. 5 and 6 show the D indices of either subject group for SER, VES and TEM 1, 2, 3 and 4 with the results of 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 the overall D index of 1.17 for SER, which retains the information of the temporal envelopes of intensity and F0. The index went 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 the similar tendency: 1.24 for SER, 0.23 for VES, 0.16 to 1.80 for TEM.

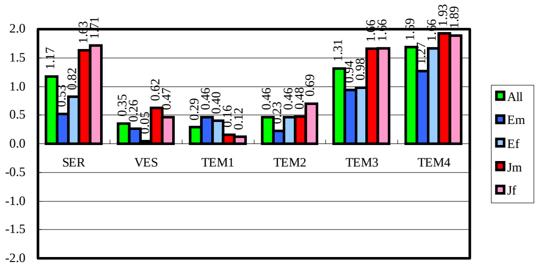

Figure 5: D indices of Japanese subjects

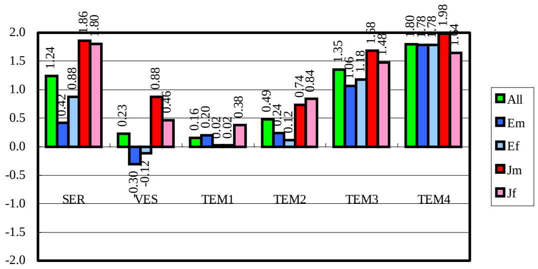

Figure 6: D indices of bilingual subjects

4. Discussion

4.1. General tendency

For both monolingual and bilingual subjects, the temporal change of intensity and pitch is not enough for LID by itself; but if other information is added, LID is possible even with a greatly degenerated signal. The TEM 1 signal holds only the intensity envelope, and the VES signal holds the intensity and pitch of vowel sections; they did not provide the subjects with sufficient information to identify the languages. In SER incomplete phonotactic information 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 the segmental information was an important contributing factor.

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. Also supporting our claim is that the scores for TEM 4 are nearly perfect (Japanese subjects: 1.69; bilingual subjects: 1.80) even though the segmental information is greatly reduced under the TEM condition.

4.2. Comparison between monolinguals and bilinguals

Bilingual subjects showed similar results to those of the Japanese subjects. Especially important is that the bilingual subjects generally registered higher D indices for Japanese stimuli than for 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 a subject's linguistic knowledge of the language, although we are not certain yet because we do not have the data from the English monolinguals.

It is also interesting that there are differences between the subject groups, as seen in Fig. 7. "Eng" and "Jap" in Fig. 7 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 segmental information, and the amount of additional information increases when it goes to the right or left. 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 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 it is thought that listeners with different linguistic knowledge use different acoustic cues for LID.

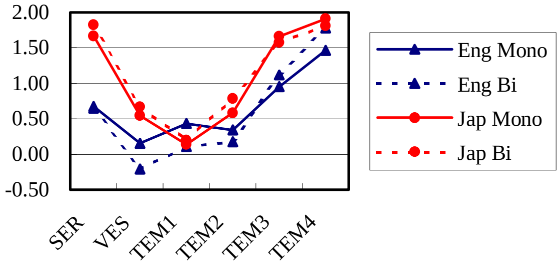

Figure 7: D indices of Japanese and bilingual subjects

5. Conclusions

Bilingual subjects showed the same general tendency as monolingual subjects. The importance of suprasegmental information combined with reduced segmental information, which we previously argued for based on only monolingual subjects, was confirmed for bilinguals.

Our experiments have also addressed the crosslinguistic difference of available cues to LID. English and Japanese are typologically different with respect to accent, rhythm and syllable structure. We suspect that such different typologies resulted in an unequal availability of prosodic cues for the two languages, which resulted in different D indices.

It has also been suggested that the listeners' linguistic knowledge affects the cues they use. Not all potential cues are perceived by all listeners. Subjects seem to have limited access to the cues according to their linguistic knowledge, as already discussed. 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.

6. References

[1] Muthusamy, Y. K., Barnard, E., and Cole, R. A., "Reviewing automatic language identification", *IEEE Signal Processing Mag.*, 11(4), 33-41, 1994.

[2] Itahashi, S., Kiuchi, T., and Yamamoto, M., "Spoken language identification utilizing fundamental frequency and cepstra", *Proc. of Eurospeech 99*, pp. 379-382, 1999.

[3] Muthusamy, Y. K., Jain, N., and Cole, R. A., "Perceptual benchmarks for automatic language identification", *Proc. of ICASSP 94*, pp. 333-336, 1994.

[4] Maidment, J. A., "Language recognition and prosody", *Speech, Hearing and Language: Work in Progress*, 1, 133-141, University College London, 1983.

[5] Ohala, J. J. and Gilbert, J. B., "Listeners' ability to identify languages by their prosody", Leon, P. and Rossi, M. (eds.) *Problèmes de Prosodie: Vol. 2, Expérimentations, Modèles et Fonctions*, pp. 123-131, Didier, Paris, 1981.

[6] Barkat, M., Ohala, J., and Pellegrino, F., "Prosody as a distinctive feature for the discrimination of Arabic dialects", *Proc. of Eurospeech 99*, pp. 395-398, 1999.

[7] Ramus, F. and Mehler, J., "Language identification with suprasegmental cues: A study based on speech resynthesis", *J. Acout. Soc. Am.*, 105, 512-521, 1999.

[8] Mori, K., Toba, N., Harada, T., Arai, T., Komatsu, M., Aoyagi, M., and Murahara, Y., "Human language identification with reduced spectral information", *Proc. of Eurospeech 99*, pp. 391-394, 1999.

[9] Muthusamy, Y. K, Cole, R. A., and Oshika, B. T., "The OGI Multi-Language Telephone Speech Corpus", *Proc. of ICSLP 92*, pp. 895-898, 1992.

[10] Lander, T., *The CSLU labeling guide*, Center for Spoken Language Understanding Technical Report, No. CSLU-014-96, Oregon Graduate Institute of Science and Technology, 1996.

[11] Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M., "Speech recognition with primarily temporal cues", *Science*, 270, 303-304, 1995.

* Kazuya Mori is currently at Tokyo Broadcasting System, Inc.

** This research was funded by Kogaku Kenkyu Shorei Enjo of The Foundation Hattori-Hokokai in 1999.