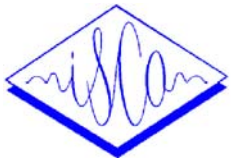

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

HUMAN LANGUAGE IDENTIFICATION WITH REDUCED SPECTRAL INFORMATION

K. Mori[1], N. Toba[1], T. Harada[1], T. Arai[1], M. Komatsu[1,2], M. Aoyagi[1], Y. Murahara[1]
[1]Sophia University, Tokyo, JAPAN
http://www.splab.ee.sophia.ac.jp/
[2]University of Alberta, Edmonton, CANADA

ABSTRACT

We conducted human language identification (LID) experiments using signals with reduced segmental information in pursuit of cues that humans use in their remarkable LID ability, which may be applicable to the development of robust automatic LID. American English and Japanese excerpts from the OGI-TS were processed by (1) spectral-envelope removal (SER) and (2) temporal-envelope modulation. With the SER signal, where the spectral-envelope is eliminated, humans could still identify the languages fairly successfully (85.2%). With the TEM signal, composed of white-noise driven, combined intensity envelopes from several frequency bands, the identification rate rose from 62.5% to 93.8% corresponding to the increasing number of bands from 1 to 4. These results, though with a limited number of languages, indicate that humans can identify languages using signal with its segmental information much reduced — in acoustic terms much reduced in spectral information.

1. INTRODUCTION

Automatic language identification (LID), a technique that recognizes which language is being spoken, is a challenging research topic; one with tremendous potential for practical applications. Considering the severe environment our speech production and perception are faced with in the real world, the development of robust automatic LID is required. As humans are quite capable of identifying languages under such conditions, we find it significant to turn to human perception ability for clues to the development of systems with a high level of reliability and optimal performance.

The sources of information that researchers can depend on for LID may be classified into two categories: segmental and non-segmental information. Much of the research so far has placed its focus on segmental information, mainly using the acoustic property of segments and their alignment ("acoustic phonetics" and "phonotactics" by the definition of Muthusamy et al. [1]). Among such research are [2], [3], [4] and [5]. On the other hand, much less attention has been directed toward the area of non-segmental information ("prosodics" by [1]).

Non-segmental information has a great deal to offer for effective LID. This is clear as we come to the realization that human speech perception is vigorously possible even when segmental information is reduced or degenerated. Obviously it is not only the segmental information but also the non-segmental, such as intonation, rhythm and stress, that humans use for speech perception. Research literature confirms this fact as in Itahashi and Du [6] and Ohala and Gilbert [7]. Itahashi and Du used temporal variation of F0 and successfully applied it to automatic LID. In the area of human LID, Ohala and Gilbert conducted perceptual experiments in 3 languages using artificial vocal pulses simulated with F0, amplitude and voice timing of human speech and confirmed that such non-segmental information provides effective cues for human LID.

We conducted human LID experiments with reduced segmental information with hopes to identify the non-segmental effect on human perception, which may be applicable to our ultimate goal of automatic LID. We used excerpts from the Oregon Graduate Institute Multi-language Telephone Speech Corpus (OGI-TS) [8], widely used for automatic LID research as well as for Muthusamy et al. [9] for human LID experiments. The use of the OGI-TS allowed our results to be made available for future comparison with other studies on both automatic and human LID. The excerpts were processed by (1) spectral-envelope removal (SER) and (2) temporal-envelope modulation (TEM) and then used within the perceptual experiments for LID. The processing methods are detailed in Section 2, and the experiments with subsequent results in Section 3. The results are discussed in Section 4.

2. SIGNAL PROCESSING

2.1 Spectral-Envelope Removal (SER)

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

The use of inverse LPC filter is based upon the concept of the AR model. Regarding the mechanism of speech generation as an AR model, LPC coefficients represent the parameters of 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.

Fig. 1 shows the 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 LPC analysis represent the impulse response of FIR filter, which performs as the inverse filter of the AR model. The output of the filter, or the residual signal, has flattened spectrum, being 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 (LPF) of 1-kHz cutoff to eliminate any spectral information that may still remain. The amplitude of the outputs of LPF was normalized among the signals using their peak values. The resultant signals were provided for the SER experiment.

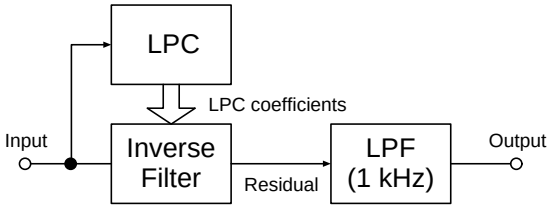

Figure 1. Block diagram of SER

10.21437/Eurospeech.1999-101

2.2 Temporal-Envelope Modulation (TEM)

In TEM, we made 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. 2 (TEM 1, 2, 3 and 4), following Shannon et al. [10].

As an illustration, Fig. 3 shows TEM 4. Speech signal was divided into 4 signals by band-pass filters, which were 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.

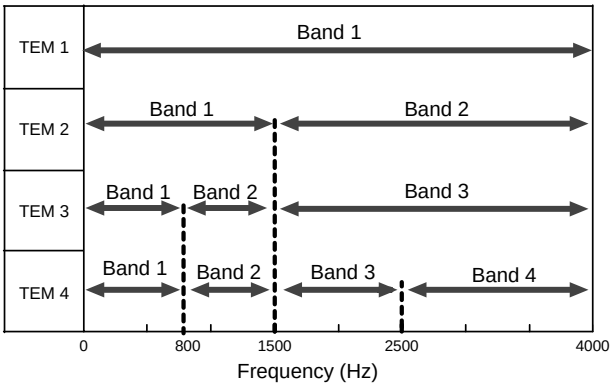

Figure 2. Frequency division of TEM

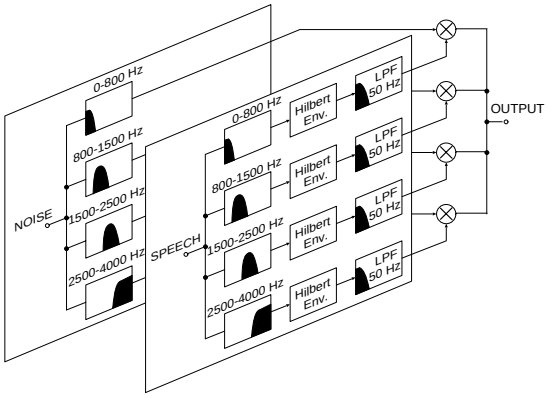

Figure 3. Block diagram of TEM 4

3. PERCEPTUAL EXPERIMENT

3.1 Extraction of original utterances

We used Japanese and English utterances in the OGI-TS [8]. The OGI-TS is a collection of telephone speech, and the utterance of each speaker in the corpus contains 1-minute spontaneous speech. We extracted two 10-second chunks of spontaneous speech from each speaker with the criteria that any parts containing excessive hesitation, pauses, proper nouns, words of foreign origin, or unnaturally foreign pronunciation should be avoided. 20 chunks each from males and females both in English and Japanese respectively were extracted, totaling 80 chunks (20 chunks * 2 sexes * 2 languages). Original utterances were thus extracted as the input for processing by SER and TEM.

3.2 Experimental stimuli

To prepare the stimuli for the SER experiment, the 80 original utterances (discussed above 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 each time for a different subject.

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 respective data set for a subject contains 80 stimuli selected out of the 320. The 80 stimuli are composed of 20 each from TEM 1, 2, 3 and 4, each TEM containing 5 each of English male/female and Japanese male/female. To preclude the subjects from learning effect, 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 and TEM experiments, 64 in total (16 * 2 sexes * 2 methods), volunteered to participate in the experiments (age: 18-29, average 21.3).

3.4 Procedure

The experiments were conducted in a sound proof chamber, using a PC. The subject used a headset to listen to the stimuli, followed the instruction on the PC display and input the responses with a mouse. First, when the subject clicked the "Play" button on the display, a stimulus was then provided through the headset. Each stimulus was presented only a single time. When the presentation finished, 4 buttons appeared: "English", "Probably English", "Probably Japanese" and "Japanese". The subject must select and click one most appropriate button according to his/her judgment of the stimulus heard. No feedback was provided. Then the "Play" button again appeared for the next stimulus. The session thus continued for 20 SER stimuli or 80 TEM stimuli at the subject's own pace. The average time required for the SER experiment was approximately 5 minutes and for the TEM experiment approximately 20 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 familiarize his/herself with the procedure. No feedback was provided for the practice session, either.

3.5 Experimental results

Table 1 shows the breakdown of actual responses given by the subjects, "E", "~E", "~J" and "J" representing "English", "Probably English", "Probably Japanese" and "Japanese" respectively. Fig. 4 shows the rates of correct identification of either language for SER 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). The results in the present study are based on the subjects' combined judgments of "E" and "~E" counted as the

judgment for the English language as well as "J" and "~J" Japanese. For SER, which retains the information of the temporal envelopes of intensity and F0, the overall identification rate was 85.2%. For TEM the overall identification rate rose from 62.5% to 93.8% as the number of bands increased from 1 to 4.

Table 1. Responses of all stimuli

Input		Response			
		E	~E	~J	J
English	SER	87	147	66	20
	TEM 1	30	189	91	10
	TEM 2	61	133	107	19
	TEM 3	144	109	43	24
	TEM 4	234	52	17	17
Japanese	SER	2	7	77	234
	TEM 1	18	121	160	21
	TEM 2	17	84	133	86
	TEM 3	8	13	38	261
	TEM 4	3	3	9	305

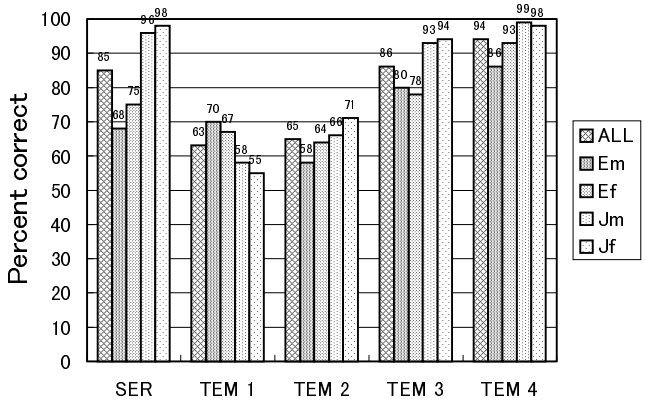

Figure 4. Identification rate for TEM and SER

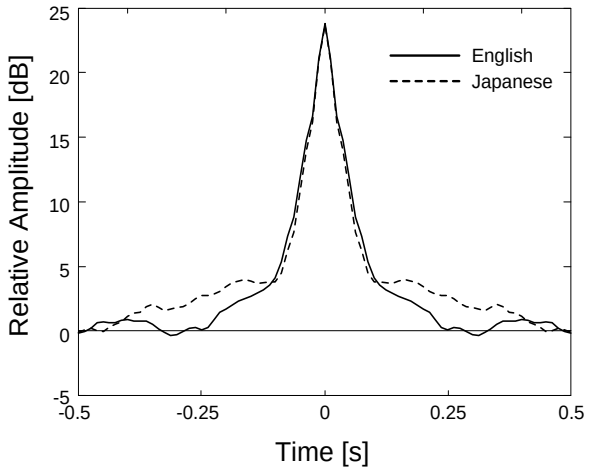

Figure 5. Impulse responses of modulation (in zero phase)

4. DISCUSSION

4.1 SER

There is a distinctive difference in the identification results between the languages but not between the sexes. The overall score for Japanese stimuli (97.2%) is higher than that for English stimuli (73.1%), which may still be seen as fairly high.

Post-experimental questionnaires indicate the subjects used the difference of 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 pitch contour information.

The subjects also answered in the questionnaires that they could often detect Japanese words in the utterances. Although the SER does not retain spectral envelope information, phonotactics is still present though not in complete form. 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. It is considered that such phonotactic information enabled the subject to spot words on occasion, and inaccurately imagined that they heard some words at some other time.

The distinctly higher scores for Japanese may have been caused by the fact that the native language of the subjects was Japanese. If we conduct the same experiments with native speakers of English, higher scores for English are anticipated. Also if we give subjects extensive training with feedback, they may grow more sensitive to the signals and be able to utilize more cues, resulting in higher scores for both languages. Thus the signals used in these experiments may embody more clues to LID than revealed at this time.

4.2 TEM

In TEM there was again a distinctive difference of scores between the languages but not between the sexes as was the case in SER. The overall scores for the Japanese stimuli (TEM 1-4: 56.6%, 68.4%, 93.4% and 98.1%) were better than their English counterparts (TEM 1-4: 68.4%, 60.6%, 79.1% and 89.4%), but the difference was not as marked as in SER. For the Japanese stimuli, the scores rise as the number of bands increases. For the English stimuli, there is not a noticeable difference between the scores 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 scores as successful results of LID, the English scores 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. Fig. 5 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 [11], English has a larger drop around 250 msec 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 of the speech recognition experiments by Shannon et al. [10]. Their results indicate that segments are more correctly identified as the number of bands increases. Our results of LID are especially

Table 2. Information contained in the signal and the results of human LID

	Segmental		Non-segmental		Results of LID
	Acoustics	Phonotactics	Intensity	Pitch	
SER	almost none	much reduced	available	available	successful
TEM 1	none		available	none	unsuccessful
TEM 2	increasing but much reduced		available	none	getting better : successful
TEM 3					
TEM 4					

similar to their result of the sentence recognition task in the respect that there is a jump from the 2-band to the 3-band conditions. For an analogy in Japanese speech, we may refer to Obata and Riquimaroux on Japanese vowels [12]. In our post-experimental questionnaires, the subjects responded that they used word spotting strategy far more often than intonational cues in the TEM experiments. The foregoing findings altogether suggest that the segmental cues are the most convincing cause of the increase in the scores from TEM 1 to 4.

5. CONCLUSION

The temporal change of intensity is not enough for LID by itself; but if other information is added, LID is quite possible even with much degenerated signal. In the experiments discussed in this paper, the TEM 1 signal holds only the intensity envelope, which did not provide the subjects with sufficient information to identify the languages. In SER pitch and phonotactic information combined with intensity information enabled better LID. In TEM 2-4 the identification rate rose as the number of bands increased, and here the segmental information was an important contributing factor. See Table 2.

In our experiments, we did not completely separate the segmental and non-segmental information as Ohala and Gilbert [7] did. Therefore we cannot conclude from our results that LID is possible solely based upon the non-segmental information. Instead, we argue that the non-segmental information — intensity and pitch — can be used under conditions where the segmental information — acoustics of segments and phonotactics — is severely reduced. This is confirmed when we see the high scores of SER (average: 85.2%), where the segments are severely degenerated but the very non-segmental information discussed above, namely intensity and pitch, are still present. Also supporting our claim is that the scores for TEM 4 are nearly perfect (average: 93.8%) though the segmental information is still much reduced in that condition.

These results, though with a limited number of languages, discovered that humans can identify languages using signal with its segmental information much reduced — in acoustic terms, much reduced in spectral information — such as signal without spectral envelope information and signal made of only a few combined frequency bands. It is our intent that such critical findings will help lead to the development or enhancement of robust and low-cost automatic LID system.

REFERENCES

[1] Muthusamy, Y.K., Barnard, E. and Cole, R.A. (1994), Reviewing automatic language identification. *IEEE Signal Processing Mag.*, vol. 11, no. 4, pp. 33-41.

[2] Muthusamy, Y.K., Berkling, K., Arai, T., Cole, R.A. and Barnard, E. (1993), A comparison of approaches to automatic language identification using telephone speech. *Proceedings of Eurospeech 93*, vol. 2, pp. 1307-1310.

[3] Arai, T. (1995), Automatic language identification using sequential information of phonemes. *IEICE Trans.*, vol. E78-D, no. 6, pp. 705-711.

[4] Yan, Y., Barnard, E. and Cole, R.A. (1996), Development of an approach to automatic language identification based on phone recognition. *Computer Speech and Language*, vol. 10, pp. 37-54.

[5] Zissman, M.A. (1996), Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31-44.

[6] Itahashi, S. and Du, L. (1995), Language identification based on speech fundamental frequency. *Proceeding of ICASSP 95*, vol. 2, pp. 1359-1362.

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

[8] Muthusamy, Y.K, Cole, R.A. and Oshika, B.T. (1992), The OGI Multi-Language Telephone Speech Corpus. *Proceedings of ICSLP 92*, vol. 2, pp. 895-898.

[9] Muthusamy, Y.K., Jain, N. and Cole, R.A. (1994), Perceptual benchmarks for automatic language identification. *Proceedings of ICASSP 94*, vol. 1, pp. 333-336.

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

[11] Arai, T. and Greenberg, S. (1997), The temporal properties of spoken Japanese are similar to those of English. *Proceedings of Eurospeech 97*, vol. 2, pp. 1011-1014.

[12] Obata, Y. and Riquimaroux, H. (1999), Role of temporal information in speech perception: Importance of amplitude envelope information. *Proceedings of the Acoustical Society of Japan, March 1999*, vol. 1, pp. 369-370 (in Japanese).