LISTENERS' ABILITY TO IDENTIFY LANGUAGES BY THEIR PROSODY

John J. Ohala and Judy B. Gilbert
University of California

Introduction

It is a common observation that different languages "sound" different not only due to their use of different segments but also due to differences in their intonation or prosody. If true, this could have considerable importance in many practical areas including second language teaching*, speech synthesis, and automatic speech recognition. Unfortunately most claims of this sort – whatever their intrinsic merit – cannot be directly applied to these areas since the claims are usually based on subjective experience and are couched in very impressionistic terms. The purpose of the research to be reported here was to attempt to verify that languages do differ in their prosody by seeing if listeners are successful in identifying the language used by a speaker when they hear only the fundamental frequency, amplitude, and certain timing characteristics of the original voice signal.

Some progress has already been made on this question. Previous investigations on this topic have used a variety of techniques and languages. To produce a signal in which segmental information has been removed, Kay Atkinson (1968) and Richard Bonte (1975) generated a pulse train having the same frequency and amplitude as the original speech signal, Clara Bush (1967) and Richardson (1973) used low-pass filtered speech,[1] and Maidment (1976) used a laryngographic signal, i.e., an indication of short-term variations in glottal electrical resistance – a signal, therefore, closely related to the original glottal waveform and uninfluenced by the resonances of the vocal tract or by supra-glottal noise sources. The languages which listeners had to identify in these studies were: English and Spanish (Atkinson), American English, British English, and Indian English (Bush),

---

* We are pleased to dedicate this paper to Professor Georges Faure who has long emphasized the necessity of including prosodic training in second language teaching.

1. Cohen and Starkweather (1959) approached the same question but it may be doubted whether their procedure of low-pass filtering speech samples at 600 Hz was sufficient to remove language-specific segmental cues.

Black English and White English (Richardson), English, French and Chinese (Bonte), and English and French (Maidment). All of these investigators used predominantly read speech. Although all investigators found better-than-chance levels of language identification, all seem to have shown that it was a difficult task since listeners' average scores could be as low as (approximately) 65% in a task involving the identification of 2 languages (Maidment) or 41% in a task involving 3 languages (Bonte). Other effects emerging from these studies are that longer stretches of speech are easier to identify than shorter stretches and that listeners correctly identify their own language better than other languages.

In the present study we sought to verify that language identification via prosody could be successfully done using conversational speech (i.e., not based on a written text) and to verify in addition, the following hypotheses:

1) long passages can be better identified than short passages.
2) the listener is more successful at making the distinction own language/not own language as opposed to the distinction other language$_1$/other language$_2$.
3) identifications will be improved if listeners are given a prior training session incorporating both the original unprocessed voice signal along with the processed version.

Due to the choice of listeners it turned out that we could also test the additional hypothesis:

4) bilingual or trilingual speakers (of the languages used in the test) outperform monolinguals.

For the languages used in the test we chose English, Japanese, and Cantonese, partly because of availability of subjects, but primarily because according to traditional prosodic descriptions they represent three distinct prosodic types: languages using stress, accent, and tone, respectively. In addition, English and Japanese are recognized as having different timing patterns: so-called "stress-timing" and "syllable-(or mora-) timing", respectively. Thus, these languages should be easily differentiated.

Method

Recordings of unrehearsed conversational speech were obtained from 12 adult males, 4 each native speakers of American English, Japanese, and Cantonese. The stimulus material for the test was drawn from the speech of 9 of these speakers (3 of each of the 3 languages); the speech

recorded from the remaining 3 speakers was used for a training sequence (see below). The recordings were obtained while each speaker, alone in a sound-treated recording room, responded to comments or questions posed him by another native speaker outside the room (whose voice the speaker could hear over an earphone). Both could see each other through a window in the recording room. From the speech recorded from each of the 9 speakers, 6 fluent passages were selected, 3 "short" passages ($\leq$10 sec) and 3 "long" passages ($\leq$15 sec, but $\geq$25 sec), yielding a total of $9 \times 6 = 54$ passages (actually, through an error the samples from one Japanese speaker included 4 short passages and only 2 long ones; thus there were a total of 28 short passages and 26 long ones). The fundamental frequency and amplitude of these passages was extracted using a custom-made circuit (Krones 1968) and fed to separate inputs of a voltage-controlled-frequency and voltage-controlled-amplitude signal generator (Wavetek Model 146) which was set to produce a triangular pulse train. When there was no fundamental frequency signal, i.e., when there was no voicing, the amplitude of the processed signal went to zero. In this way the original speech signal was converted to a "buzz" having the same frequency, amplitude, and timing (at least with respect to the relative timing of voice on/voice off).

In addition, a training passage was constructed which included speech samples from the 3 extra speakers, both in their original unprocessed form and in the processed "buzz" form. The training passage plus instructions followed by the 54 test samples in "buzz" form, randomized, were dubbed onto a master stimulus tape for presentation to listeners. Each test item was followed by a short tone marking the end of the item plus 5 seconds of silence during which the language was to be identified. The instructions, which were presented in writing as well as on the tape, were as follows:

> This is a test to find out if people can recognize a language based on the intonation pattern alone. That is, if everything is removed but the "music" of the speech, can the language be identified? An example of intonation patterns in English is the following: 1. 'You're not going?' 2. 'You're not going!' Nothing was changed in those sentences but the intonation pattern.
>
> This test consists of recorded excerpts taken from conversations in Japanese, Cantonese, and English. These excerpts have been electronically processed so that only the musical part, the intonation, remains. The "music" representation sounds like a buzz. We will ask you to listen to this buzz and guess what language it is from. Half of the items are short and half are long [however, see above]. Each item will be followed by a marker sound, like this [tone] and then you will have a 5 second pause to mark the answer sheet. If you think the buzz was from English, please write 'E' after the

item number. If you think the buzz was from Japanese, write 'J'; Cantonese, write 'C'. There are many speakers of the same language included in the samples. The items are in random order.

To help you hear the buzz tone better, we are going to give you a brief training period. You are going to hear a phrase taken from a sentence in English. It will be followed by a buzz representing that phrase. [Samples played].

Now you will hear the entire sentence, including the phrase, followed by the buzz representation. [Samples played.]

Now you will hear a sample of Cantonese, followed by the buzz representation. [Samples played.]

Next you will hear a sample of Japanese, followed by the buzz representation. [Samples played.]

We want to be sure that you understand the procedure for this test, so we are going to give you some test samples. Please mark your answer sheet 'C' if you think the buzz was Cantonese; 'J' if you think it was Japanese; or 'E' if you think it was English. [Three short samples played.]

You are now ready to begin the test. None of the voices used for this test were included in the examples you heard before.

The stimulus tape was presented individually to 41 adult listeners over a high-quality playback system via earphones. Of the 41 listeners, 18 were native speakers of American English, 12 of Cantonese, and 11 of Japanese. It was not feasible to be completely selective with regards to listeners' mastery of or familiarity with the other 2 languages used in the test aside from their native language. As it turned out there were 12 monolinguals, all native English speakers, 25 "bilinguals" (including those who had studied one of the other languages extensively in school), and 4 trilinguals, all among the native Cantonese speakers. The training passage was deliberately withheld from 5 English speaking listeners (4 monolinguals and 1 bilingual); all other listeners heard the training passage. The answer sheet consisted of 56 numbered squares of which only 54 were used.

Results

The results are given in Tables 1 through 12. 'E,' 'J,' and 'C' stand for 'English,' 'Japanese,' and 'Cantonese', respectively.

Listed on the vertical axis of Tables 1 through 10 (i.e., the labels for the rows) are the languages as presented; listed on the horizontal axis (i.e., labels for the columns) are the languages given as responses by the listeners.

Thus in Table 1, for example, it can be seen that the monolingual listeners identified Japanese sentences as English 51 times. With the exception of Table 11, the numbers along the diagonal running from top left to bottom right represent correct responses.

Table 1. Monolinguals (N = 12)

Presented \ Heard	E	J	C
E	140	51	25
J	52	85	79
C	46	69	101

correct: 50.3%

Table 2. Bilinguals (N = 25)

	E	J	C
E	280	113	57
J	87	238	125
C	85	76	289

correct: 59.3%

Table 3. Trilinguals (N = 4)

	E	J	C
E	34	13	25
J	22	40	10
C	23	8	41

correct: 53.2%

Table 4. English (with training passage) (N = 13)

	E	J	C
E	168	39	27
J	50	102	82
C	53	64	117

correct: 55.1%

Table 5. English (no training) (N = 5)

	E	J	C
E	49	29	12
J	26	31	33
C	19	32	39

correct: 44.1%

Table 6. Japanese (N = 11)

	E	J	C
E	107	74	17
J	39	95	64
C	43	31	124

correct: 54.9%

Table 7. Cantonese (N = 12)

	E	J	C
E	130	35	51
J	46	135	35
C	39	26	151

correct: 64.2%

Table 8. Short Passages

	E	J	C
E	222	92	55
J	83	201	126
C	98	61	210

correct: 55.1%

Table 9. Long Passages

	E	J	C
E	232	85	52
J	78	162	88
C	56	92	221

correct: 57.7%

Table 10. Total (all conditions all listeners)

	E	J	C
E	454	177	107
J	161	363	214
C	154	153	431

correct: 56.4%

Table 11. Correct Identifications

	Listeners' Own Lang.	Other Languages
Observed	463	785
Expected by Chance	416	832

Table 12. Misidentifications

	Other Lang. 's as Listeners' own	Other lang.$_1$ as other lang.$_2$ (and vice-versa)
Observed	339	352
Expected by Chance	345.5	345.5

Discussion of Results

The overall score (number correct identifications/total responses) was 56.4% (or 58.1% if we exclude the responses of those 5 listeners who were not given the training session). This is far above the chance level of 33.3% and is therefore highly significant statistically ($p < 0.001$ via one-tailed Chi-square test on the 1×2 matrix consisting of correct vs. incorrect responses). The scores of those English listeners who heard the training session was higher than those who did not, 55.1% as opposed to 44.1% and this difference is highly significant. The scores for monolinguals, bilinguals, and trilinguals were 50.3%, 59.3%, and 53.2%, respectively. Of these, only the bilinguals' score is significantly better than the others. Long passages were better identified than short passages, 57.7% versus 55.1%; this difference does achieve statistical significance ($p<0.05$). Listeners'

identification of their own language is better than identification of other languages by a greater proportion than the 1:2 ratio which would be expected by chance (although this trend is reversed among the Japanese listeners) ($p << 0.01$). However, although there are slightly fewer misidentifications of other languages as listeners' own language than there are misidentifications of other $language_1$ as other $language_2$ (and vice-versa), as was predicted by hypothesis (2) above, (cf. Table 12), this difference is far from achieving statistical significance and, in any case, as a trend is only true of the English listeners.

General Discussion

These results are in general agreement with previous studies and extend their validity to free conversation. Although the scores were somewhat higher than previous comparable studies, probably due to the training passage, the clear prosodic differences between the languages, and, possibly, the fact that free conversation was used (thus preserving more of the "natural rhythm" of the languages), the scores were still low enough to justify characterizing the task as difficult. It is possible that the method of converting the speech signal into a "buzz" destroys certain crucial prosodic information, e.g., syllable or word boundaries, which may be necessary to evaluate any language-specific aspects of fundamental frequency variation which are synchronized with those units. Based on these promising results, however, and our own experience with this technique, we would propose that future work in this area approach the following topics:

1) The relative importance in this task of the various prosodic parameters present in the "buzz" versions of speech, i.e., fundamental frequency, amplitude, and timing, could be evaluated by holding most of them constant while one is allowed to vary—a relatively easy thing to do if a voltage-controlled-frequency and voltage-controlled-amplitude signal generator is used, as in the present study. (Elimination of all timing information, of course, would be difficult.)

2) Would it be possible to raise listeners' scores even more by giving them a more extensive and more elaborate training session or, possibly, by giving them feedback on the correctness of their identifications during the course of the test? If so, this might be a way of getting second language learners' "attuned" to the specific prosodic patterns of the language they are seeking to learn.

3) If such a test is performed on a large number of languages and listeners' responses presented as confusion matrices it may be possible to verify traditional linguistic prosodic classifications of languages as "syllable-timed" or "stress-timed" on the one hand and "tone language", "accent language", etc., on the other by seeing which languages are confused most often with which others under these circumstances.

Acknowledgements

We gratefully acknowledge the technical assistance of Steve Pearson, Karen Bracco, and Catherine Rodriguez-Nieto and the support of the Language Laboratory, University of California, Berkeley, and the National Science Foundation (via a grant to the Phonology Laboratory).

BIBLIOGRAPHY

ATKINSON, K. (1968) Language identification from non-segmental cues. *Working Papers in Phonetics* (UCLA) 10, 85-89.

BONTE, R. (1975) *Can you identify a language by its prosody?* Unpub. M.A. Thesis, University of California, Berkeley.

BUSH, C. (1967) Some acoustic parameters of speech and their relationships to the perception of dialect differences. *TESOL Quarterly* 1, 20-30.

COHEN, A. and STARKWETHER, J. (1961) Vocal cues to the identification of language. *American J. Psychology* 74,90-93.

KRONES, R. (1968) Calibrating the pitch extractor. *Monthly Internal Memorandum* (Berkeley), May, 39-42.

MAIDMENT, J. (1976) Voice fundamental frequency characteristic as language differentiators. *Speech and Hearing Work in Progress*, University College, London 2, 74-93.

RICHARDSON, J.A.C. (1973) *The identification by voice of speakers belonging to two ethnic groups*. Unpub. Ph. D. Dissertation, Ohio State University.