Reviewing Human Language Identification

Masahiko Komatsu

School of Psychological Science, Health Sciences University of Hokkaido
Ainosato 2-5, Sapporo, 002-8072 Japan
koma2@hoku-iryo-u.ac.jp

Abstract. This article overviews human language identification (LID) experiments, especially focusing on the modification methods of stimulus, mentioning the experimental designs and languages used. A variety of signals to represent prosody have been used as stimuli in perceptual experiments: lowpass-filtered speech, laryngograph output, triangular pulse trains or sinusoidal signals, LPC-resynthesized or residual signals, white-noise driven signals, resynthesized signals preserving or degrading broad phonotactics, syllabic rhythm, or intonation, and parameterized source component of speech signal. Although all of these experiments showed that "prosody" plays a role in LID, the stimuli differ from each other in the amount of information they carry. The article discusses the acoustic natures of these signals and some theoretical backgrounds, featuring the correspondence of the source, in terms of the source-filter theory, to prosody, from a linguistic perspective. It also reviews LID experiments using unmodified speech, research into infants, dialectology and sociophonetic research, and research into foreign accent.

Keywords: Language identification, Human language identification, Speech modification, Source-filter model, Prosody.

1 Introduction

Language Identification (LID) is a process for identifying a language used in speech.[1] Although there have been several reviews of automatic LID by computers ([3][4][5] etc.), there have been no extensive reviews of human, or perceptual, LID research as far as the author knows. As opposed to the well-documented automatic LID research, the research scene of human LID gives the impression that it is not well traffic-controlled and the studies are often sporadic. The backgrounds and motivations of researchers are diverse. Thus, the research into the human capability of LID extends into several disciplines, and the communication seems lacking between disciplines, sometimes even within a discipline.

The cues for identifying languages are classified into two types: segmental and prosodic. The former includes "acoustic phonetics," "phonotactics," and "vocabulary," and the latter corresponds to "prosodics" of the terms in [3]. In the field of automatic LID by computers, much of the research so far has focused on utilizing

[1] Part of this article is based on [1][2].

C. Müller (Ed.): Speaker Classification II, LNAI 4441, pp. 206–228, 2007.

segmental features contained in the speech signal, although some research also suggests the importance of incorporating prosodic information into the system ([6][7] etc.). In contrast to this engineering research scene, most of the research on perceptual LID by humans has focused on prosodic information.

Humans' capacity for LID has drawn the attention of engineers, linguists, and psychologists since 1960s. The typical method of research is to conduct perceptual experiments with stimulus signals that are supposed to contain prosodic information of certain languages but not contain segmental information. In other words, the signals are used as the representative of prosody. The modification methods of stimulus signals and the languages used in the experiments have been various and not consistent across researchers. The critical question here is whether the signals used really represent the prosody of language, or more specifically what represents prosody acoustically.

This article aims at giving the reader an overview of the human LID research, discussing the modification methods of speech, the experimental designs, and the relations to the prosodic types of used languages. It also introduces examples from related areas of research. The latter part of the article discusses the acoustic correlates of prosody to advance suggestions for future research.

2 Overview of Human LID Experiments

2.1 LID with Modified Speech

A variety of signals and languages have been used as stimuli in perceptual experiments (see Table A1). All studies listed there have used modified speech that was presumed to represent the prosody of speech, and all of them have concluded that prosody plays some role in LID.

Of the stimuli to represent prosody used in previous experiments, the handiest is lowpass-filtered speech. Atkinson [8] used this signal for the discrimination test of English and Spanish, and showed that these two languages were discriminable and that error rates varied depending on speech styles. The lowpass-filtering technique is still being used (e.g., Mugitani et al. [9], for Eastern and Western Japanese, which have different characteristics of lexical accent).

The most straightforward is a laryngograph signal, which is an indication of variations in glottal electrical resistance, closely related to the glottal waveform. It sounds like a dull buzzing noise, varying in pitch. Maidment [10][11] showed that English and French are discriminable with this signal. Moftah & Roach [12] compared the lowpass-filtered and laryngograph signals and concluded that there was no significant difference in language identification accuracy for Arabic and English.

A synthesized signal was used by Ohala and Gilbert [13]. They made triangular pulse trains that had the same F0 and amplitude as the original speech signal; the amplitude was set to zero where F0 was unavailable, i.e., there was no voicing. The signal simulated the F0, amplitude, and voice timing of the original speech, and sounded like a buzz. They designed the experiments to investigate the relation of prosodic types of languages to explicitly defined acoustic features. They chose three prosodically different languages to test: English (stress-accented, stress-timed), Japanese (pitch-accented, mora-timed), and Chinese (tonal). The results indicated that these languages were discriminated. It also showed that the listeners with prior training performed better than those with no training, that bilingual listeners performed better than trilinguals and monolinguals, and that longer samples were better discriminated than shorter ones. Barkat et al. [14] used sinusoidal signals instead of triangular pulses to test Western and Eastern Arabic, the former of which loses short vowels causing prosodic difference. These two dialects were discriminated by Arabic listeners, but not by non-Arabic listeners.

Application of Linear Predictive Coding (LPC) is comparatively new in the history of research on human LID. LPC separates the speech signal into the source and filter, or spectral, components in terms of the source-filter model. The idea of using LPC can be traced back to Foil's experiment [15], but it was simply a preparatory test for developing an automatic LID system. Foil resynthesized speech by LPC with its filter coefficients constant, resulting in the speech signal that had a constant spectrum all the time, and said that languages were easy to discriminate with this signal. The languages discriminated were not explicitly described.

Navrátil [16] used an inverse LPC filter to remove spectral information of speech; the signal represented prosody of speech. He also made a random-spliced signal, where short segments roughly corresponding to syllables were manually cut out and concatenated in a random order; the resultant signal lost F0 and intensity contours of the original speech and represented syllable-level phonotactic-acoustic information plus duration. He compared the LID results with these signals for Chinese, English, French, German, and Japanese, and concluded that prosody contributes less to LID (see Table 1).

Table 1. Correct identification rates for 6-s excerpts in Navrátil's experiment [16] (Chance level: 20%). Random-spliced speech represnts syllable information, and inverse-LPC-filtered speech represents prosody.

Stimulus	German	English	French	Japanese	Chinese	Overall
Unmodified speech	98.7 %	100.0 %	98.7 %	81.7 %	88.7 %	96.0 %
Random-spliced	79.7 %	98.7 %	79.1 %	54.6 %	57.7 %	73.9 %
Inverse-LPC-filtered	32.1 %	34.3 %	69.4 %	45.3 %	65.9 %	49.4 %

Komatsu et al. [17] used an inverse LPC filter, and further lowpass-filtered the signal with the cutoff of 1 kHz to ensure spectral removal. The resultant signal sounded like muffled speech. They suspected that partial phonotactic information remained in this signal, so they also created the consonant-suppressed signal for comparison, where the amplitude of consonant intervals of the former signal was set to zero to remove possible consonantal effects. In the former signal, the LID for English and Japanese was successful; but in the latter consonant-suppressed signal, it was unsuccessful. Besides, they created signals driven by band-limited white noise. These signals were the replication of what Shannon et al. [18] used for speech recognition experiments. The speech was divided into 1, 2, 3, or 4 frequency bands, the intensity contours of these bands were used to modulate noises of the same bandwidths, and they were summed up altogether. The resultant signals kept only intensity when the number of bands was 1, and broad spectral information increased as the number of bands increased. The correct identification rate increased as the number of bands increased. Comparing the results with all these stimulus types (see Fig. 1), they concluded that LID was possible using signals with segmental information drastically reduced; it was not possible with F0 and intensity only, but possible if partial phonotactic information was also available. The results also suggested the variation due to prosodic difference of languages and listeners' knowledge.

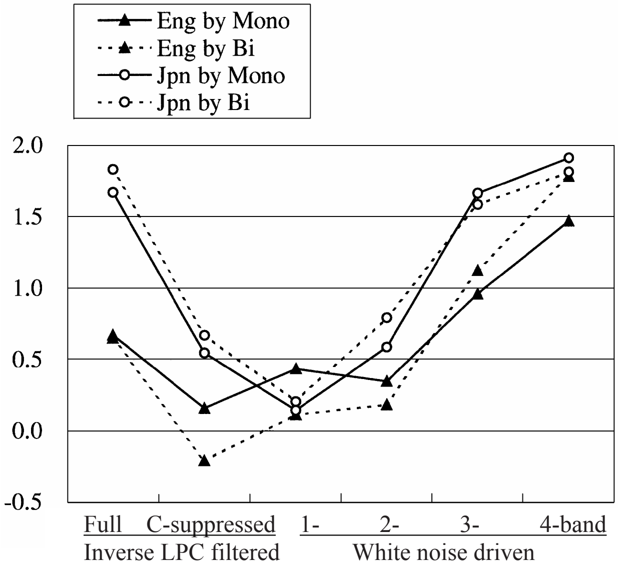

Fig. 1. LID results of English and Japanese in terms of the discriminability index by Komatsu et al. [17]. The index was calculated such that "English" and "Japanese" were scored as +/−2 and "Probably English" and "Probably Japanese" were +/−1, where positive values indicate correct responses and negative, incorrect ones. The graph indicates the results of each stimulus type for English and Japanese samples identified by Japanese monolingual listeners and Japanese-English bilingual listeners, respectively. C-suppressed inverse-LPC-filtered and 1-band white-noise-driven stimuli have only prosodic information (F0, intensity), and the amount of additional information increases when it goes to either side of the graph.

The idea of using LPC was taken a step further by Komatsu et al. [19]. They decomposed the source signal, in terms of the source-filter model, into three parameters, F0, intensity, and Harmonics-to-Noise Ratio (HNR); and created stimulus signals simulating some or all parameters from white noise and/or pulse train. Compared to the previous LPC applications, this method has the merits of the parameterization of the source features and the completeness of spectral removal. They conducted a perceptual discrimination test using excerpts from Chinese, English, Spanish, and Japanese, differing in lexical accent types and rhythm types. In general, the results indicated that humans can discriminate these prosodic types and that the discrimination is easier if more acoustic information is available (see Fig. 2). Further, the results showed that languages with similar rhythm types are difficult to discriminate (i.e., Chinese-English, English-Spanish, and Spanish-Japanese). As to accent types, tonal/non-tonal contrast was easy to detect. They also conducted a preliminary acoustic analysis of the experimental stimuli and found that quick F0 fluctuations in Chinese contribute to the perceptual discrimination of tonal and non-tonal. However, their experiment had a drawback that the number of experimental conditions was too large, which as a consequence had the number of repetitions in each condition too small to run statistical tests. Experiments must be designed to zero in on fewer combinations of acoustic parameters and languages in future.

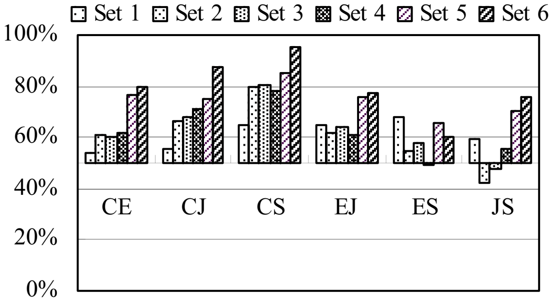

Fig. 2. Correct response rates for each language pair by Komatsu et al. [19]. “C” stands for Chinese, “E” for English, “J” for Japanese, and “S” for Spanish. Set 1 is stimuli made of white noise simulating the intensity of the original speech, Set 2 is made of pulse train simulating intensity, Set 3 is made of white noise and pulse train simulating intensity and HNR, Set 4 is made of pulse train simulating F0, Set 5 is made of pulse train simulating intensity and F0, and Set 6 is made of white noise and pulse train simulating intensity, HNR, and F0. Sets 1-3 represent amplitude-related information, Set 4 represents F0 information, and Sets 5-6 represent both information.

The modification method by Ramus and Mehler [20] is different from others; they are segment-based. They conducted perceptual experiments on English and Japanese, controlling broad phonotactics, syllabic rhythm, and intonation. They segmented the original English and Japanese speech into phonemes and replaced them by French phonemes to exclude the segmental cues to LID. They created four types of stimulus signals differing in the information they contain: “saltanaj”, “sasasa”, “aaaa”, and “flat sasasa”. In “saltanaj”, all fricatives were replaced by /s/, stops by /t/, liquids by /l/, nasals by /n/, glides by /j/, and vowels by /a/. In “sasasa”, all consonants were replaced by /s/, and vowels by /a/. In “aaaa”, all segments were replaced by /a/. “Flat sasasa” was the same as “sasasa” but its F0 was made constant. The information that each stimulus contained and the results of LID tests are summarized in Table 2. Ramus and Mehler concluded that syllabic rhythm is a necessary and sufficient cue.

Table 2. Stimuli and LID results of Ramus and Mehler [20]. "+" indicates presence of cue, and "–" indicates absence of cue.

	Intonation	Syllabic rhythm	Broad phonotactics	Result of LID
saltanaj	+	+	+	successful
sasasa	+	+	–	successful
aaaa	+	–	–	unsuccessful
flat sasasa	–	+	–	successful

Although all of these experiments showed that "prosody" plays some role in LID, the stimuli used differ from each other in the amount of information they carry; that is, the acoustic definitions of prosody are not coherent among the studies. An appropriate selection of stimuli is needed for further research.

Experimental procedures in these studies are simple. Participants were provided with a stimulus and instructed to identify a language or dialect. Many experiments simply adopted a multiple choice from two or more language names. Others used somewhat different procedures. In Atkinson's experiment [8], the ABX procedure was used. Ramus and Mehler [20] used a multiple choice from two fictional language names. Maidment [11] and Komatsu et al. [17] used the 4-point scale judgment, e.g., definitely French, probably French, probably English, and definitely English; and Mugitani et al. [9] used the 5-point scale judgment. Komatsu et al. [19] asked the sequential order of the presented stimuli because a multiple choice from four languages would be so difficult to discourage the participants: e.g., participants listened to a Chinese sample and an English sample sequentially and judged whether it was Chinese-English or English-Chinese.

Experimental designs started with a simple one. Discrimination tests were performed for a pair of popular languages: English and Spanish (Atkinson [8]), and English and French (Maidment [10][11]). Mugitani et al. [9] was a pretest for an infants' experiment. Moftah and Roach [12] intended to compare the previously used signals using Arabic and English. Ohala and Gilbert [13] designed their experiment to investigate the relation of prosodic types of languages to explicitly defined acoustic features. They chose three prosodically different languages to test, English (stress-accented, stress-timed), Japanese (pitch-accented, mora-timed), and Chinese (tonal), as well as exploring several other effects. They used conversational speech while preceding studies had predominantly used reading. Barkat et al. [14] focused on the prosodic difference between two Arabic dialects caused by short vowel elision. Navrátil's experiment [16] intended to compare the contributions of prosodic and segmental features, covering five languages. Komatsu et al. [17] compared the LID with segmental features reduced by several methods using English and Japanese. Komatsu et al. [19] parameterized the source features and involved four languages differing in prosodic types (stress-accented, pitch-accented, tonal; stress-timed, syllable-timed, mora-timed) to discuss the relation of the acoustic features to prosodic types. Ramus and Mehler [20] focused on the rhythmic difference of English (stress-timed) and Japanese (mora-timed), which backed up their argument on the acoustic correlates of rhythm.

2.2 LID with Unmodified Speech

It should also be noted that some researchers have used real speech as the stimulus. The purposes and methods of these experiments are different from those using modified speech (see Table A2 for details).

An engineering motivation is the benchmark by humans. Muthusamy et al. [21] did this using 1-, 2-, 4-, and 6-s excerpts of spontaneous speech of 10 languages. The listeners were given feedback on every trial. The obtained results showed that humans are quite capable of identifying languages, but the perceptual cues were not experimentally explored. The cues were sought by Navrátil [16] using two types of modified speech as well as unmodified one, mentioned in section 2.1.

Barkat and Vasilescu [22] sought perceptual cues by two experiments. One is a dialect identification of six Arabic dialects. Endogenous listeners were better at identifying dialects than exogenous listeners. The other used the AB procedure for five Romance languages. The perceptual space was configured by Multi-Dimensional Scaling (MDS) with familiarity and vowel system configuration.

Maddieson and Vasilescu [23] conducted experiments with five languages, combining identification and similarity judgment, and showed that individual variation is poorly explained by prior exposure to the target languages and academic linguistic training.

Bond et al. [24] explored the features that listeners attend using 11 languages from Europe, Asia, and Africa. They used magnitude estimation and MDS techniques and showed that languages were deployed by familiarity, speaker affect (reading dramatic or not), and prosodic pattern (rhythm and F0).

Stockmal et al. [25] challenged to remove the effects of speakers' identity. They did experiments with the AB procedure and similarity judgment for several language pairs using the speech samples spoken by the same bilingual speakers. The results indicated that the listeners discriminated the language pairs spoken by the same speakers and that, in the MDS configuration, they used rhythm information within the context of language familiarity. Stockmal and Bond [26] further eliminated the effect of language familiarity. They replicated the previous experiment only with languages unfamiliar to listeners. The selected languages were all African: all of them are syllable-timed, and all but Swahili were tonal. The results suggested that the listeners discriminated the language pairs using difference in the phoneme inventories.

2.3 Examples from Other Related Areas of Research

Experiments have been conducted with somewhat different perspectives, too. Table A3 listed a few examples of research into infants. Boysson-Bardies et al. [27] showed that the babbling of 8-month-old infants is discriminable by adults. Non-segmental cues such as phonation, F0 contour, and intensity were important. Hayashi et al. [28] and Mugitani et al. [9][29] indicated that infants can discriminate their native language or dialect from others. They used the head-turn preference procedure, which regards the stimulus that infants pay attention longer as preferred, and showed that infants paid attention to their native language or dialect for a longer duration. The original interest of Ramus and Mehler [20], who did the experiment with adults, is in exploring how pre-language infants discriminate languages in bilingual or trilingual environments. See [20][30][31][32] for more literature.

Perceptual experiments have been conducted also for dialectology and sociophonetic purposes (see Table A4 for several recent examples). They seek the perceptual cues of dialect identification and measure the distance among dialects.

Van Bezooijen and Gooskens [33] compared the identification rates between the original speech and monotonized (F0 flattened) speech, representing segmental features, or lowpass-filtered speech, representing prosody, for four Dutch dialects. The results indicated that prosody plays a minor role in dialect identification. A follow-up experiment using only the unmodified signal showed that the difference in the identification rates between spontaneous speech and reading varies among dialects. They also conducted the experiment for five British English dialects, showing that prosody plays a minor role as in Dutch dialects. Gooskens and van Bezooijen [34] adopted a different procedure, 10-point scale judgment of whether dialectal or standard, for six Dutch dialects and six British English dialects. They showed that segmentals are more important, as in their previous experiments, and that the importance of prosody is somewhat larger in English than in Dutch. Gooskens [35] explored 15 Norwegian dialects, and showed that endogenous listeners identify dialects better than exogenous listeners and that prosody is more important than in Dutch dialect identification.

In the United States, Thomas and Reaser [36] did a discrimination test of English spoken by African Americans and European Americans. In order to focus on phonetic characteristics, speech samples were carefully selected to include diagnostic vowels, usually /o/, and subject pronouns, related to intonation variation, but to avoid diagnostic morphosyntactic and lexical variables. European American listeners performed better with monotonized samples than with lowpass-filtered samples; and the detailed analysis indicated that African Americans could not use the vowel quality as a perceptual cue. Thomas et al. [37], who incorporated different techniques, converting all vowels to schwa and swapping F0 and segmental durations, showed that the vowel quality is important although F0 also plays a role and that different listener groups use different cues.

See [36][38] for extensive reviews of the studies in these areas, including experiments with various modification techniques: lowpass-filtering, highpass-filtering, center-clipping; lowpass-filtering vs. monotonization (F0 flattening); bandwidth compression to remove nasality; backward playing, temporal compression; F0 level change of isolated vowels; F2 modification to make vowels front or back; resynthesis of /s/-/ʃ/ to assess the McGurk effect on the perceptual boundary; a synthetic vowel continuum; synthetic vowels; synthetic diphthongs; modification of the intonation and the speaking rate; unmodified, lowpass-filtered, random-spliced, vs. written text.

Table A5 gives examples of the research into foreign accents.[2] Miura et al. [40] and Ohyama and Miura [41] did experiments manipulating a segmental feature (PARCOR[3] coefficients) and prosodic features (F0, intensity, phoneme durations),

[2] See also [39].

[3] PARCOR stands for partial auto-correlation.

showing prosodic features contribute more. Miwa and Nakagawa [42] focused on only prosodic features and showed that the sensitivity to such a feature is different between native and non-native listeners.

A confounding factor of perceptual experiments on LID or the naturalness of languages is that prosody is closely related to not only linguistic information but also paralinguistic and nonlinguistic information. Grover et al. [43] found that F0 variation at the continuation junctures of English, French, and German differ significantly, but that the synthetically replaced intonation patterns were regarded by listeners as speakers' variation of emotional attitudes or social classes rather than foreign accents.

Another problem was raised by Munro [44], who investigated the effect of prosody on the perception of foreign accent using lowpass-filtered speech. The results indicated that the foreign-accentedness was recognized in the lowpass-filtered speech. However, they did not show a correlation with the unfiltered, or original, speech, which means that samples regarded as accented when lowpass-filtered may not be regarded as accented when not filtered, suggesting that listeners may use different cues in different conditions.

3 Acoustic Definition of Prosody

3.1 Reviewing Stimulus Signals

The speech modification methods described in section 2.1 may be classified into several groups. The first one is what does not use synthesis or resynthesis techniques: lowpass-filtering and laryngograph output. The second one, which uses synthesis/resynthesis techniques, includes the simple acoustic simulation (triangular pulses, sinusoidal signals, band-limited white noise) and the signal processing based on the source-filter model (inverse LPC filtering, source feature parameterization). Random splicing and phoneme replacing constitute the third group: these modify the signal in segment-based manners, permuting or replacing them, rather than utilizing acoustic processing globally. The second group may be called more "acoustic," and the third group more "phonological."

It is in question whether some of them do properly represent prosody in speech. In lowpass filtering, the cutoff frequency is usually set at 300-600 Hz to make speech unintelligible, but it is reported that speech is sometimes intelligible if repeatedly listened to [45]. In lowpass-filtered speech, some segmental information is preserved under the cutoff frequency, F0 sometimes rises higher than the cutoff, and intensity is not preserved [20]. A perceptual experiment confirmed that, if the cutoff is set at 300 Hz, the filtered signal retains prosodic features and some laryngeal voice quality features but not articulatory features [45]. The laryngograph output is an indication of short-term variations of glottal electrical resistance and virtually uninfluenced by supraglottal resonance and noise source [12][13]. This means that it is not representative of output speech, which we actually hear in usual situations. Due to the loss of resonance and noise source, it does not contain sonority information, which will be discussed in section 3.2. The simple acoustic simulation techniques are close to the source-filter-model-based ones but incomplete because they lack something. The simulation of prosody with pulse or sinusoidal trains does not take the noise source into account. The white-noise driven signal keeps the intensity contour of the original speech but does not have any other information such as F0.

Of the segment-based approaches, random-splicing, of course, destroys prosodic contour information as the experimenter intends. It was reported that speech random-spliced with the segment size between 150-300 ms was unintelligible, and a perceptual experiment confirmed that speech random-spliced with the segment size of 255 ms carries voice quality, some articulatory features, and overall prosodic features (level, range, and variability of pitch, loudness, and sonority) but loses tempo [45]. In Navrátil's experiment [16], segments in length roughly corresponding to syllables were manually cut out.

The processing based on the source-filter model may be the best to represent prosody (see the discussion in section 3.2), but it may have a technical drawback. Inverse LPC filtering does not guarantee the perfect removal of the spectrum. To avoid this problem, Komatsu et al. [17] used a lowpass filter in conjunction with an inverse LPC filter, but still reported that some listeners said they spotted words although it is not clear whether it was true or illusory. On the other hand, the source feature parameterization, in which the stimulus is made of pulses and white noise from scratch, is perfect in the spectral removal but problematic especially in the F0 contour estimation. Komatsu et al. [19] used the MOMEL algorithm [46], originally devised to extract the intonation contour of the intonation languages (i.e., non-pitch-accented, non-tonal). It seems that the algorithm does not only remove microprosody but affects the F0 variation related to pitch accent and tone [47].

To estimate F0 correctly and compare among languages, a model that does not incorporate any phonology of specific languages is desired. For example, modeling by INTSINT [46], which simply encodes F0 patterns, seems more adequate for the present purpose than ToBI [48][49], which describes only F0 variations meaningful in respective languages. Another desirable nature of the model is to divide the contour into components. Although the difference in F0 between languages of different prosodic types have been pointed out [50], local characteristics seem more important than global characteristics [6][51]. Further, three types of F0 characteristics varying across languages have been distinguished: global, recurrent, and local [52]. Although there have been proposed various F0 models [53], not all are adequate for the present purpose. Scrutiny of models is necessary for future research.

The notion of rhythm is also confounding. Since Pike's dichotomy of stress- and syllable-timed rhythms [54], the isochronic recurrence of stress/syllable in speech signal has not been found. This has caused the definition of rhythm to be claimed variously [31][55]. Timing hypotheses argue that there is an isochronic unit or that the length of the higher level unit such as a word can be predictable from the number of lower level units. Rhythm hypotheses argue that the rhythmic difference is the reflection of structural factors, such as syllable structures, phonotactics, etc., rather than timing specifically [30][56][57]. There are also other alternative claims focusing the competence of coordinating units in speech production [58], or the role of the unit in perception [59].

Ramus and Mehler's experiment [20] was to support the rhythm hypothesis. They define "syllabic rhythm" as the temporal alignments of consonant and vowel, which is the reflection of syllable structures, and showed that it is essential to the perceptual discrimination of languages. Here, rhythm is not defined by acoustic features such as the intensity contour but defined by the discrimination of consonant and vowel. This reminds us of the role of broad phonotactics in human LID [17] and automatic LID [60].

The studies pursuant to the rhythm hypothesis are rather phonological than acoustic, because they need phoneme identification. Phonemes must be identified in the stream of speech signal prior to the measurement of durations. However, it seems that this method has been taken as an expedient because appropriate acoustic measures to grasp syllable shapes were not available. Ramus et al. [30] (p. 271 fn) states that "[their] hypothesis should ultimately be formulated in more general terms, e.g. in terms of highs and lows in a universal sonority curve." In another analysis [57], devoiced vowels are treated as consonantal rather than vocalic to reflect more acoustic features. Retesting this hypothesis by calculating sonority in acoustic terms [61] is worth mentioning.

3.2 Correspondence Between Acoustic and Linguistic Features

This section argues that the source at the acoustic level approximately represents prosody at the linguistic level. Fig. 3 shows the simplified correspondence of the articulatory, acoustic, and linguistic models. Note that the figure is simplified for illustration, and that the correspondences of the features in different models are actually not as simple as drawn in the figure. When humans utter speech, especially vowels, the voice source is created at the larynx, is modulated by the vocal tract, and results in the speech sound. This can be modeled by an acoustic model called the source-filter model, in which the source, or the excitation signal, is processed by the filter, resulting in the speech signal. The source consists of three physical elements, F0, intensity, and HNR; and the filter determines the spectral envelope of the sound in the frequency domain. Very naively, prosodic, or suprasegmental, features in the linguistic model seem to involve F0 and intensity of the speech signal controlled by the laryngeal activity: the tone and accent systems seem to involve F0 and intensity, and rhythm seems to involve the temporal variation of intensity. On the other hand, segmental features, i.e. phoneme distinctions, seem to be related to spectral patterns determined by the vocal tract shape, or the movement of articulators. However, their correspondence to each other is actually not so simple. For example, in the recognition of phonemes, it is known that various acoustic cues interact, including not only the spectral pattern but also F0 and intensity. So far, the acoustic contributors to prosodic features have not been thoroughly inquired into. This section discusses whether, or how well, the source elements of the acoustic model approximately represent the linguistic prosody.

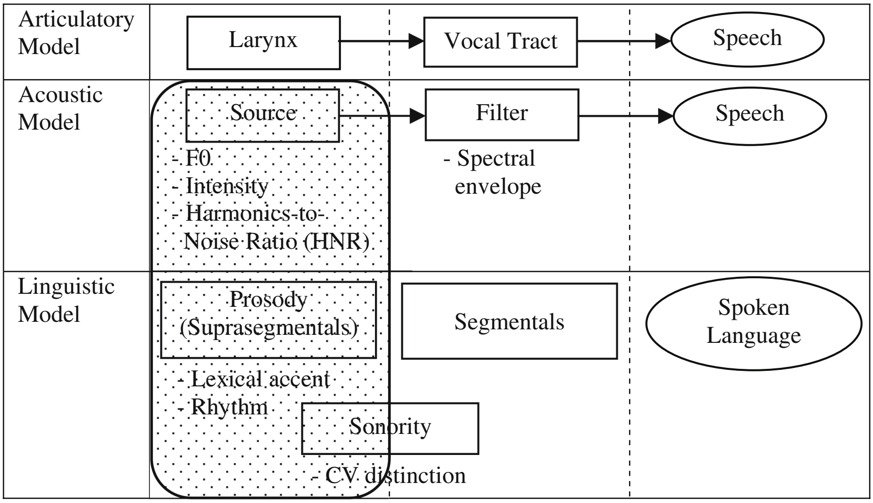

Fig. 3. Approximate correspondence of articulatory, acoustic, and linguistic models. The shaded area indicates the correspondence discussed in this section.

Linguistic features that constitute a prosodic typology include lexical accent (stress accent, pitch accent, and tone), intonation, and rhythm (stress-timed, syllable-timed, and mora-timed). Their acoustic correlates are, basically, F0, intensity, and length. However, assuming that rhythm is, even if partly, the reflection of syllable structures, it follows that acoustic properties that represent sonority contribute to constituting rhythm.

Sonority is a linguistic feature that approximately represents syllable shapes (see Sonority Sequencing Principle [62]). The sonority feature is ambivalently prosodic and segmental by nature. On one hand, it represents syllable shapes, and consequently contributes to rhythm. On the other hand, it is closely related to the articulatory manner of segments, and, as a result, it partially represents some phoneme classes and phonotactics. Consequently, the acoustic properties that represent sonority contain both prosodic and segmental information. Then, it is impossible to completely separate acoustic features corresponding to prosody from acoustic features corresponding to segmentals. The dichotomy of prosody and segmentals are possible in the linguistic model but impossible in the real-world acoustic model.

The important question is, therefore, whether or how the source features at the acoustic level represent sonority, and do not represent segmental features, at the linguistic level. To this end, experiments on Japanese consonant perception were conducted with the LPC residual signal [63]. The identification rate of major classes corresponding to the sonority ranks, i.e., obstruent, nasal, liquid, and glide [62], was as high as 66.4 % while that of phonemes was as low as 20.0 % (chance level: 1/17 = 5.9 %).

Further, to investigate how sonority is represented in the source, the confusion matrix obtained from this experiment was analyzed with MDS [64]. The analysis showed that sonority can be located in a multi-dimensional perceptual space, and that the dimensions of the space have correspondence to both acoustic and phonological features. Because the LPC residual signals represent the source, the confusion pattern for the signals indicates the consonants' similarities in the source. Although fitting of the data was not satisfactory, the result showed that the consonants can be modeled in a 3-dimensional perceptual space according to their sonority ranks. Its dimensions could be related to acoustic measurements and phonological features. The result also showed that sonority can be mostly defined within the source.

In the perceptual space, consonants with the same sonority rank clearly tended to cluster together. As seen in Fig. 4, voiceless plosives, voiceless fricatives, voiced obstruents, and nasals/glides gathered together. Each dimension of the perceptual space had correspondence to acoustic and phonological features, as shown in Table 3. The dimensions were correlated with acoustic measurements obtained from the stimulus, and had correspondence to some of the sonority-related distinctive features [65].

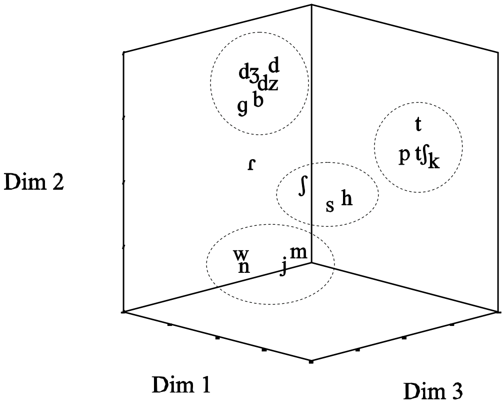

Fig. 4. Three-dimensional analysis of consonant perception in the LPC residual signal (altered from [64])

Table 3. Correspondence of each dimension to acoustic and phonological features

		Lower sonority ⟵⟶	Higher sonority
Dim 1	Acoustics	Lower HNR	Higher HNR
	Phonology	[–voice]	[+voice]
Dim 2	Acoustics	Smaller amplitude Lower F0	Larger amplitude Higher F0
	Phonology	[–sonorant]	[+sonorant]
Dim 3	Acoustics	Shorter duration	Longer duration
	Phonology	[–continuant]	[+continuant]

These results indicate that the source retains sonority information while segmental information, such as cues for phoneme identification, is effectively suppressed. The importance of sonority, or broad phonotactics, has been shown by many previous LID studies, human or automatic.

4 Concluding Remarks

This article started with overviewing human LID experiments, especially focusing on the modification methods of stimulus, also mentioning the experimental designs and languages used (section 2). It was followed by the discussion on what those acoustic features used in human LID experiments mean (section 3). It discussed the acoustic natures of the stimulus signals and some theoretical backgrounds, featuring the correspondence of the source to prosody.

LID is, from a linguistic point of view, a study on the naturalness of a language and the difference from other languages. Language is defined as the pair of the form and meaning. LID research focuses on the form only, and would provide cross-linguistic foundations for the description of the form. Simple manipulations of acoustic features may suffice to engineering purposes; their linguistic meanings have not been inquired into. The author hopes that this article gives the reader some insights into this question.

Acknowledgments. The author is grateful to Dennis R. Preston, Charlotte Gooskens, Erik R. Thomas, and Jiří Navrátil for their kind offer of information.

References

1. Komatsu, M.: What constitutes acoustic evidence of prosody? The use of Linear Predictive Coding residual signal in perceptual language identification. LACUS Forum 28, 277–286 (2002)
2. Komatsu, M.: Acoustic constituents of prosodic types. Doctoral dissertation. Sophia University, Tokyo (2006)
3. Muthusamy, Y.K., Barnard, E., Cole, R.A.: Reviewing automatic language identification. IEEE Signal Processing Magazine 11(4), 33–41 (1994)
4. Zissman, M.A., Berkling, K.M.: Automatic language identification. Speech Communication 35, 115–124 (2001)
5. Navrátil, J.: Automatic language identification. In: Schultz, T., Kirchhoff, K. (eds.) Multilingual speech processing, pp. 233–272. Elsevier, Amsterdam (2006)
6. Thymé-Gobbel, A.E., Hutchins, S.E.: On using prosodic cues in automatic language identification. In: Proceedings of International Conference on Spoken Language Processing '96, pp. 1768–1771 (1996)
7. Itahashi, S., Kiuchi, T., Yamamoto, M.: Spoken language identification utilizing fundamental frequency and cepstra. In: Proceedings of Eurospeech '99, pp. 383–386 (1999)
8. Atkinson, K.: Language identification from nonsegmental cues [Abstract]. Journal of the Acoustical Society of America 44, 378 (1968)
9. Mugitani, R., Hayashi, A., Kiritani, S.: Developmental change of 5 to 8-month-old infants' preferential listening response. Journal of the Phonetic Society of Japan 4(2), 62–71 (2000) (In Japanese)
10. Maidment, J.A.: Voice fundamental frequency characteristics as language differentiators. Speech and Hearing: Work in Progress 2. University College, London, pp. 74–93 (1976)

11. Maidment, J.A.: Language recognition and prosody: Further evidence. Speech, Hearing and Language: Work in Progress 1. University College, London, pp. 133–141 (1983)
12. Moftah, A., Roach, P.: Language recognition from distorted speech: Comparison of techniques. Journal of the International Phonetic Association 18, 50–52 (1988)
13. Ohala, J.J., Gilbert, J.B.: Listeners' ability to identify languages by their prosody. In: Léon, P., Rossi, M. (eds.) Problèmes de prosodie: Expérimentations, modèles et fonctions. Didier, Paris, vol. 2, pp. 123-131 (1979)
14. Barkat, M., Ohala, J., Pellegrino, F.: Prosody as a distinctive feature for the discrimination of Arabic dialects. In: Proceedings of Eurospeech '99, pp. 395–398 (1999)
15. Foil, J.T.: Language identification using noisy speech. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing '86, pp. 861–864 (1986)
16. Navrátil, J.: Spoken language recognition: A step toward multilinguality in speech processing. IEEE Transactions on Speech and Audio Processing 9, 678–685 (2001)
17. Komatsu, M., Mori, K., Arai, T., Aoyagi, M., Murahara, Y.: Human language identification with reduced segmental information. Acoustical Science and Technology 23, 143–153 (2002)
18. Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M.: Speech recognition with primarily temporal cues. Science 270, 303–304 (1995)
19. Komatsu, M., Arai, T., Sugawara, T.: Perceptual discrimination of prosodic types and their preliminary acoustic analysis. In: Proceedings of Interspeech 2004, pp. 3045–3048 (2004)
20. Ramus, F., Mehler, J.: Language identification with suprasegmental cues: A study based on speech resynthesis. Journal of the Acoustical Society of America 105, 512–521 (1999)
21. Muthusamy, Y.K., Jain, N., Cole, R.A.: Perceptual benchmarks for automatic language identification. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing '94, pp. 333–336 (1994)
22. Barkat, M., Vasilescu, I.: From perceptual designs to linguistic typology and automatic language identification: Overview and perspectives. In: Proceeding of Eurospeech 2001, pp. 1065–1068 (2001)
23. Maddieson, I., Vasilescu, I.: Factors in human language identification. In: Proceedings of International Conference on Spoken Language Processing 2002, pp. 85–88 (2002)
24. Bond, Z.S., Fucci, D., Stockmal, V., McColl, D.: Multi-dimensional scaling of listener responses to complex auditory stimuli. In: Proceedings of International Conference on Spoken Language Processing '98, vol. 2, pp. 93–95 (1998)
25. Stockmal, V., Moates, D.R., Bond, Z.S.: Same talker, different language. In: Proceedings of International Conference on Spoken Language Processing '98, vol. 2, pp. 97–100 (1998)
26. Stockmal, V., Bond, Z.S.: Same talker, different language: A replication. In: Proceedings of International Conference on Spoken Language Processing 2002, pp. 77–80 (2002)
27. Boysson-Bardies, B., de Sagart, L., Durand, C.: Discernible differences in the babbling of infants according to target language. Journal of Child Language 11, 1–15 (1984)
28. Hayashi, A., Deguchi, T., Kiritani, S.: Reponse patterns to speech stimuli in the headturn preference procedure for 4- to 11-month-old infants. Japan Journal of Logopedics and Phoniatrics 37, 317–323 (1996)
29. Mugitani, R., Hayashi, A., Kiritani, S.: The possible preferential cues of infants' response toward their native dialects evidenced by a behavioral experiment and acoustical analysis. Journal of the Phonetic Society of Japan 6(2), 66–74 (2002)
30. Ramus, F., Nespor, M., Mehler, J.: Correlates of linguistic rhythm in the speech signal. Cognition 73, 265–292 (1999)

31. Tajima, K.: Speech rhythm and its relation to issues in phonetics and cognitive science. Journal of the Phonetic Society of Japan 6(2), 42–55 (2002)
32. Hayashi, A.: Perception and acquisition of rhythmic units by infants. Journal of the Phonetic Society of Japan 7(2), 29–34 (2003) (In Japanese)
33. van Bezooijen, R., Gooskens, C.: Identification of language varieties: The contribution of different linguistic levels. Journal of Language and Social Psychology 18, 31–48 (1999)
34. Gooskens, C., van Bezooijen, R.: The role of prosodic and verbal aspects of speech in the perceived divergence of Dutch and English language varieties. In: Berns, J., van Marle, J. (eds.) Present-day dialectology: Problems and findings. Mouton de Gruyter, Berlin, pp. 173–192 (2002)
35. Gooskens, C.: How well can Norwegians identify their dialects? Nordic Journal of Linguistics 28, 37–60 (2005)
36. Thomas, E.R., Reaser, J.: Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. Journal of Sociolinguistics 8, 54–87 (2004)
37. Thomas, E.R., Lass, N.J., Carpenter, J.: Identification of African American speech. In: Preston, D.R., Niedzielski, N. (eds.) Reader in Sociophonetics. Cambridge University Press, Cambridge (in press)
38. Thomas, E.R.: Sociophonetic applications of speech perception experiments. American Speech 77, 115–147 (2002)
39. Gut, U.: Foreign accent. In: Müller, C. (ed.) Speaker classification. LNCS, vol. 4343, pp.75–87, Springer, Heidelberg (2007)
40. Miura, I., Ohyama, G., Suzuki, H.: A study of the prosody of Japanese English using synthesized speech. In: Proceedings of the 1989 Autumn Meeting of the Acoustical Society of Japan, pp. 239–240 (1989) (In Japanese)
41. Ohyama, G., Miura, I.: A study on prosody of Japanese spoken by foreigners. In: Proceedings of the 1990 Spring Meeting of the Acoustical Society of Japan, pp. 263–264 (1990) (In Japanese)
42. Miwa, T., Nakagawa, S.: A comparison between prosodic features of English spoken by Japanese and by Americans. In: Proceedings of the 2001 Autumn Meeting of the Acoustical Society of Japan, pp. 229-230 (2001) (In Japanese)
43. Grover, C., Jamieson, D.G., Dobrovolsky, M.B.: Intonation in English, French and German: Perception and production. Language and Speech 30, 277–295 (1987)
44. Munro, M.J.: Nonsegmental factors in foreign accent: Ratings of filtered speech. Studies in Second Language Acquisition 17, 17–34 (1995)
45. van Bezooijen, R., Boves, L.: The effects of low-pass filtering and random splicing on the perception of speech. Journal of Psycholinguistic Research 15, 403–417 (1986)
46. Hirst, D., Di Cristo, A., Espesser, R.: Levels of representation and levels of analysis for the description of intonation systems. In: Horne, M. (ed.) Prosody: Theory and experiment, pp. 51–87. Kluwer Academic, Dordrecht, The Netherlands (2000)
47. Komatsu, M., Arai, T., Sugawara, T.: Perceptual discrimination of prosodic types. In: Proceedings of Speech Prosody 2004, pp. 725–728 (2004)
48. Venditti, J.J.: Japanese ToBI labelling guidelines. Manuscript, Ohio State University, Columbus (1995)
49. Pierrehumbert, J.: Tonal elements and their alignment. In: Horne, M. (ed.) Prosody: Theory and experiment, pp. 11–36. Kluwer Academic, Dordrecht, The Netherlands (2000)
50. Eady, S.J.: Differences in the F0 patterns of speech: Tone language versus stress language. Language and Speech 25, 29–42 (1982)

51. Komatsu, M., Arai, T.: Acoustic realization of prosodic types: Constructing average syllables. LACUS Forum 29, 259–269 (2003)
52. Hirst, D., Di Cristo, A.: A survey of intonation systems. In: Hirst, D., Di Cristo, A. (eds.) Intonation systems: A survey of twenty languages, pp. 1–44. Cambridge University Press, Cambridge (1998)
53. Shih, C., Kochanski, G.: Prosody and prosodic models. In: Tutorial at International Conference on Spoken Language Processing 2002, Denver CO (2002)
54. Pike, K.L.: The intonation of American English. University of Michigan Press, Ann Arbor (1945)
55. Warner, N., Arai, T.: Japanese mora-timing: A review. Phonetica 58, 1–25 (2001)
56. Dauer, R.M.: Stress-timing and syllable-timing reanalyzed. Journal of Phonetics 11, 51–62 (1983)
57. Grabe, E., Low, E.L.: Durational variability in speech and the Rhythm Class Hypothesis. In: Gussenhoven, C., Warner, N. (eds.) Laboratory phonology 7. Mouton de Gruyter, Berlin, pp. 515–546 (2002)
58. Tajima, K.: Speech rhythm in English and Japanese: Experiments in speech cycling. Doctoral dissertation, Indiana University, Bloomington, IN (1998)
59. Cutler, A., Otake, T.: Contrastive studies of spoken-language perception. Journal of the Phonetic Society of Japan 1(3), 4–13 (1997)
60. Nakagawa, S., Seino, T., Ueda, Y.: Spoken language identification by Ergodic HMMs and its state sequences. IEICE Transactions J77-A(2), 182–189 (1994) (In Japanese)
61. Galves, A., Garcia, J., Duarte, D., Galves, C.: Sonority as a basis for rhythmic class discrimination. In: Proceedings of Speech Prosody 2002, pp. 323–326 (2002)
62. Clements, G.N.: The role of the sonority cycle in core syllabification. In: Beckman, M.E., Kingston, J. (eds.) Papers in laboratory phonology 1, pp. 283–333. Cambridge University Press, Cambridge (1990)
63. Komatsu, M., Tokuma, W., Tokuma, S., Arai, T.: The effect of reduced spectral information on Japanese consonant perception: Comparison between L1 and L2 listeners. In: Proceedings of International Conference on Spoken Language Processing 2000, vol. 3, pp. 750–753 (2000)
64. Komatsu, M., Tokuma, S., Tokuma, W., Arai, T.: Multi-dimensional analysis of sonority: Perception, acoustics, and phonology. In: Proceedings of International Conference on Spoken Language Processing 2002, pp. 2293–2296 (2002)
65. Blevins, J.: The syllable in phonological theory. In: Goldsmith, J.A. (ed.) The handbook of phonological theory, pp. 206–244. Basil Blackwell, Cambridge, MA (1995)

Appendix: Lists of LID Research

Table A1. LID using modified speech

Atkinson (1968) [8]

Language:	English, Spanish
Material:	Poetry, prose, natural speech, nursery rhymes, dramatic dialogues
Modification:	Lowpass-filetered
Method:	Identification (ABX)
Result:	English and Spanish were discriminated. Least error rates in poetry, greatest in prose and nursury rhymes.

Table A1. (*continued*)

Mugitani, Hayashi, & Kiritani (2000) [9]
Language: Eastern Japanese dialect, Western Japanese dialect
Material: Elicited speech by a speaker fluent in both dialects
Modification: (1) Unmodified, (2) Lowpass-filtered (400Hz)
Method: (1) 5-pt scale (+2=Definitely Eastern, -2=Never Eastern; +2=Definitely Western, -2=Never Western), (2) Identification (Eastern or not)
Result: (1) Almost perfect, (2) Significant result

Maidment (1976) [10]
Language: English, French
Material: Reading
Modification: Laryngograph waveform
Method: Identification
Result: 64.5%

Maidment (1983) [11]
Language: English, French
Material: Spontaneous speech
Modification: Laryngograph waveform
Method: 4-pt scale judgment (1=Definitely French, 4=Definitely English)
Result: 74.68% [Calculated such that both "1 Definitely French" and "2 Probably French" counted as French and both "3 Definitely English" and "4 Probably English" counted as English]

Moftah & Roach (1988) [12]
Language: Arabic, English
Material: Reading and spontaneous speech
Modification: (1) Laryngograph waveform, (2) Lowpass-filtered (500Hz)
Method: Identification
Result: (1) 63.7%, (2) 65.5%

Ohala & Gilbert (1979) [13]
Language: English, Japanese, Cantonese Chinese
Material: Conversation
Modification: Triangle pulses simulating F0, amplitude, voice timing
Method: Identification
Result: 56.4% [Chance level: 33.3%]

Barkat, Ohala, & Pellegrino (1999) [14]
Language: Western Arabic dialects, Eastern Arabic dialects
Material: Elicited story-telling
Modification: (1) Unmodified, (2) Sinusoidal pulses simulating F0, amplitude, voice timing
Method: Identification
Result: (1) 97% by Arabic listeners, 56% by non-Arabic listeners. (2) 58% by Arabic listeners, 49% by non-Arabic listeners

Table A1. (*continued*)

Foil (1986) [15]
Language: Unknown (one Slavic and one tonal SouthEast Asian languages?)
Material: Unknown (noisy radio signals?)
Modification: LPC-resynthesized with the filter coefficients constant
Method: Identification
Result: Easy to distinguish

Navrátil (2001) [16]
Language: Chinese, English, French, German, Japanese
Material: Spontaneous speech?
Modification: (1) Unmodifed, (2) Random-splicing, (3) Inverse-LPC-filtered
Method: Identification
Result: (1) 96%, (2) 73.9%, (3) 49.4% [Chance level: 20%]

Komatsu, Mori, Arai, Aoyagi, & Murahara (2002) [17]
Language: English, Japanese
Material: Spontaneous speech
Modification: (1) Inverse-LPC-filtered followed by lowpass-filtered (1kHz), (2) Consonant intervals of (1) suppressed, (3) Band-devided white-noise driven (from 1 to 4 bands)
Method: 4-pt scale judgment (1=English, 4=Japanese)
Result: For English, (1) 70.0%, (2) 44.0%, (3) 56.0-95.0% varying over the number of bands; for Japanese, (1) 100.0%, (2) 66.0%, (3) 60.0-96.0% varying over the number of bands

Komatsu, Arai, & Sugawara (2004) [19]
Language: Chinese, English, Japanese, Spanish
Material: Reading
Modification: (1) White noise simulating intensity, (2) Pulse train simulating intensity, (3) Mixture of white noise and pulse train simulating intensity and harmonicity, (4) Pulse train simulating F0, (5) Pulse train simulating intensity and F0, (6) Mixture of white noise and pulse train simulating intensity, harmonicity, and F0
Method: Judgment on the sequential order by listening to a language pair
Result: (1) 61.3%, (2) 61.1%, (3) 63.1%, (4) 62.8%, (5) 74.7%, (6) 79.3% [Chance level: 50%]

Ramus & Mehler (1999) [20]
Language: English, Japanese
Material: Reading
Modification: Resynthesized preserving (1) broad phonotactics, rhythm, and intonation, (2) rhythm and intonation, (3) intonation only, (4) rhythm only
Method: Identification (using fictional language names)
Result: (1) 66.9%, (2) 65.0%, (3) 50.9%, (4) 68.1%; indicating the importance of rhythm

Table A2. LID using unmodified speech only

Muthusamy, Jain, & Cole (1994) [21]	
Language:	10 languages (English, Farsi, French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, Vietnamese)
Material:	Spontaneous speech
Method:	Identification
Result:	With 6-s excerpts, 69.4% (varying from 39.2 to 100.0% over languages) [Chance level: 10%]
Barkat & Vasilescu (2001) [22]	
Language:	6 Arabic dialects
Material:	Elicited speech
Method:	Identification
Result:	78% for Western dialicts, 32% for Eastern dialects by Western Arabic listerns; 59% for Western dialects, 90% for Eastern dialects by Eastern Arabic listeners [Chance level: 16.7%]
Language:	5 Romance languages (French, Italian, Spanish, Portuguese, Romanian)
Material:	Reading or story-telling
Method:	AB (same or different)
Result:	MDS configured with familiarity, vowel system complexity
Maddieson & Vasilescu (2002) [23]	
Language:	5 languages (Amharic, Romanian, Korean, Morroccan Arabic, Hindi)
Material:	Reading
Method:	(1) Identification, (2) Identification and similarity judgment
Result:	(1) 65% [Chance level: 20%], (2) Partial identification patterns varied among languages
Bond, Fucci, Stockmal, & McColl (1998) [24]	
Language:	11 languages from Europe, Asia, Africa (Akan, Arabic, Chinese, English, French, German, Hebrew, Japanese, Latvian, Russian, Swahili)
Material:	Reading
Method:	Similarity to English (magnitude estimation)
Result:	MDS configured with familiarity, speaker affect, prosodic pattern (rhythm, F0)
Stockmal, Moates, & Bond (1998) [25]	
Language:	Language pairs (Arabic-French, Hebrew-German, Akan-Swahili, Latvian-Russian, Korean-Japanese, Ombawa-French, Ilocano-Tagalog)
Material:	Each language pair was spoken by the same talker
Method:	(1) AB (same or different), (2) AB (7-pt similarity; 1=very dissimilar, 7=very similar)
Result:	(1) 66.5% and 63.4% depending on the experimental condition [Chance level: 50%], (2) 5.19 for the same-language pairs, 3.45 for the different-language pairs
Stockmal & Bond (2002) [26]	
Language:	Language pairs (Akan-Swahili, Haya-Swahili, Kikuyu-Swahili, Luhya-Swahili)
Material:	Reading; each language pair was spoken by the same talker
Method:	AB (same or different)
Result:	71% [Chance level: 50%]

Table A3. Examples of research into infants

Boysson-Bardies, Sagart, & Durand (1984) [27]		
	Language:	Arabic, Chinese, French
	Material:	8- and 10-month-old infants' babbling
	Method:	Choice of French from French-Arabic or French-Chinsese pair
	Result:	For 8- and 10-month samples respectively; 75.8%, 74.4% (French-Arabic pairs); 69.4%, 31.9% (French-Chinese pairs) [Chance level: 50%]
	Language:	Arabic, French
	Material:	6-, 8-, 10-month-old infants' babbling
	Method:	Choice of French from the pair of babbling
	Result:	For 6-, 8-, 10-month samples respectively; 55.5-68%, 67.5-74%, 49-56.9% (varying over experimental conditions) [Chance level: 50%]
Hayashi, Deguchi, & Kiritani (1996) [28]		
	Language:	Japanese, English
	Material:	Spontaneous speech by a bilingual speaker
	Method:	Head-tern preference procedure (for infants)
	Result:	Infants aged over 200 days preferred the native language Japanese
Mugitani, Hayashi, & Kiritani (2000) [9]		
	Language:	Eastern Japanese dialect, Western Japanese dialect
	Material:	Elicited speech by a speaker fluent in both dialects
	Modification:	Unmodified
	Method:	Head-turn preference procedure (for infants)
	Result:	Greater preference to their native Eastern dialect
Mugitani, Hayashi, & Kiritani (2002) [29]		
	Language:	Eastern Japanese dialect, Western Japanese dialect
	Material:	Elicited speech by a speaker fluent in both dialects
	Modification:	Lowpass-filtered (400Hz)
	Method:	Head-turn preference procedure (for infants)
	Result:	8-month-old infants preferred their native Eastern dialect

Table A4. Examples of dialectology and sociophonetic research using modified speech

Van Bezooijen & Gooskens (1999) [33]		
	Language:	4 Dutch dialects
	Material:	Spontaneous
	Modification:	(1) Unmodified, (2) Monotonized (flat f0), (3) Lowpass-filtered (350Hz)
	Method:	Identification of Country, Region, Province, and Place
	Result:	(1) Country 90%, Region 60%, Province 40%; (2) Decreased from (1) by 7%, 2%, 4%; (3) Decreased from (1) by 29%, 41%, 32% [Chance levels for Country, Region, Province are 50%, 12.5%, 5.26% respectively; there were almost no answer for Place]
	Language:	5 British English dialects
	Material:	Spontaneous
	Modification:	(1) Unmodified, (2) Monotonized (flat f0), (3) Lowpass-filtered (350Hz), (4) The same as (3) but including typical dialect prosody

Table A4. (*continued*)

Method:	Identification of Country, Region, Area, and Place
Result:	(1) Country 92%, Region 88%, Area 52%; (2) Decreased from (1) by 4%, 10%, 3%; (3) Decreased from (1) by 18%, 43%, 33%, (4) Increased from (3) by 5%, 5%, 1% [Chance levels for Country, Region, Area are 50%, 14.28%, 6.67% respectively; there were almost no answer for Place]
Gooskens & van Bezooijen (2002) [34]	
Language:	6 Dutch dialects
Material:	Interview
Modification:	(1) Unmodified, (2) Monotonized (flat F0), (3) Lowpass-filtered (350Hz)
Method:	10-pt scale judgment (1=dialect, 10=standard)
Result:	(1)(2) 4 groups separated, (3) Standard variation and the others were separated
Language:	6 British English dialects
Material:	Interview
Modification:	(1) Unmodified, (2) Monotonized (flat F0), (3) Lowpass-filtered (350Hz)
Method:	10-pt scale judgment (1=dialect, 10=standard)
Result:	(1)(2) 3 groups separated, (3) 2 groups separated
Gooskens (2005) [35]	
Language:	15 Norwegian dialects
Material:	Reading
Modification:	(1) Unmodified, (2) Monotonized (flat F0)
Method:	Identification (marking on a map, choosing from 19 countries), Similiarity to the listener's own dialect (15-pt scale)
Result:	(1) 67% by endogenous listeners, 25% by exogenous listeners, (2) 50% by endogenous listeners, 16% by exogenous listeners [Chance level: 5.3%]
Thomas & Reaser (2004) [36]	
Language:	English spoken by African Americans and European Americans
Material:	Spontaneous speech (interview)
Modification:	(1) Unmodified, (2) Monotonized (flat F0), (3) Lowpass-filtered (330Hz)
Method:	Identification
Result:	(1) 71.10%, (2) 72.08%, (3) 52.28% by European American listeners [Chance level: 50%]
Thomas, Lass, & Carpenter (in press) [37]	
Language:	English spoken by African Americans and European Americans
Material:	Reading
Modification:	(1) Unmodified, (2) Monotonized (flat F0), (3) Conversion of all vowels to schwa
Method:	Identification
Result:	Vowel quality is important; F0 also plays a role
Language:	English spoken by African Americans and European Americans
Material:	Reading
Modification:	Swapping F0 and segmental durations
Method:	Different listener groups use different cues

Table A5. Examples of research into foreign accent using modified speech

Miura, Ohyama, & Suzuki (1989) [40]	
Language:	English spoken by Japanese speakers
Material:	Reading
Modification:	Substitution of the features of a native speaker's English with those of Japanese English (PARCOR coefficients, F0, Intensity, Phoneme durations)
Method:	Choosing the more natural sample from a pair of samples
Result:	Durations and F0 contribute to the naturalness
Ohyama & Miura (1990) [41]	
Language:	Japanese spoken by English, French, Chinese speakers
Material:	Reading
Modification:	Substitution of the features of foreign-accented Japanese with those of a native speaker's Japanese (PARCOR coefficients, F0, Intensity, Phoneme durations)
Method:	Choosing the more natural sample from a pair of samples
Result:	Durations contribute for the speech by English and French speakers; F0 contribute for the speech by Chinese speakers
Miwa & Nakagawa (2001) [42]	
Language:	English spoken by native speakers and Japanese
Material:	Reading
Modification:	Resynthesized preserving (1) F0 and intensity, (2) F0, (3) intensity
Method:	Judgment on naturalness (5-pt scale)
Result:	English spoken by native speakers were more natural. Japanese instructors of English were less sensitive to intensity variation than native instructors
Grover, Jamieson, & Dobrovolsky (1987) [43]	
Language:	English, French, German
Material:	Reading
Modification:	Replacement of the continuative pattern of F0 with that of another language
Method:	Choosing the more natural sample from a pair of samples
Result:	Not discriminated
Munro (1995) [44]	
Language:	English spoken by Mandarin Chinese and Canadian English speakers
Material:	(1) Elicited sentence, (2) Spontaneous speech
Modification:	Lowpass-filtered (225Hz for male speech, 300Hz for female speech)
Method:	Judgment on accentedness (1=Definitely spoken with a foreign accent, 4=Definitely spoken by a native speaker of English)
Result:	For Mandarine speakers (1) 1.8, (2) 2.1; for Canadian English speakers (1) 3.0, (2) 2.8