ON USING PROSODIC CUES IN AUTOMATIC LANGUAGE IDENTIFICATION

Ann E. Thymé-Gobbel and Sandra E. Hutchins

Natural Speech Technologies, Inc.

ABSTRACT

This paper presents an effort to explore the utility of prosodic information in language identification/ discrimination (LID) tasks. We present our model and results from pair-wise LID tasks with English, Spanish, Japanese and Mandarin using multi-speaker elicited spontaneous speech and a selected set of prosodic parameters. These languages represent four different types of languages, varying in pitch use and timing. Parameters were designed to capture pitch and amplitude contours on a syllable-by-syllable basis, and to be insensitive to overall amplitude, pitch, and speaking rate.

Results show that prosodic cues alone can distinguish between some language pairs with results comparable to many non-prosodic systems, indicating that prosodic parameters are highly useful in automatic LID. However, the statistical relationships between a number of individual features deduced from timing and pitch measurements are needed to begin to capture such complex perceptual events as rhythm. Strengths of individual prosodic parameters and classes of parameters –primarily pitch, secondarily duration and location – reflect differences between the four languages mostly as expectated based on the linguistic literature, suggesting that effective use of prosodic parameters is aided by an understanding of the relationships between physical measurements and perceived linguistic events.

1. INTRODUCTION

Although language identification/ discrimination (LID) has been researched for the past 2 decades, a current renewed interest can be linked to the establishment of the Oregon Graduate Institute Multi-language Telephone Speech Corpus (OGI-TS), described in [6]. The utility of accurate LID includes the ability of automatically tailoring a speech-based tool, such as online banking or information retrieval, to the native language of the user.

LID approaches and results are summarized in [4] and [9]. Past approaches include HMM and NN models, expert systems and various clustering algorithms which have used raw waveforms, broad phonetic features, detailed acoustic features, formant vectors, pitch contours, vocabulary, and so on. Until recently, few systems have included prosodic measures. Earlier systems that did consider prosodic measures found such measures marginally successful. Muthusamy [5] found that speech rate and syllable timing offered little to improve results in his system. Savic et al. [7] considered pitch change over the duration of the sentence and the word, and found tendencies toward differences between tone languages and Indo-European languages.

Our continuing effort explores the utility of prosodic information in LID tasks. Prosodic cues include stress, rhythm, and intonation. Each cue is a complex, language dependent perceptual entity expressed primarily as a combination of three physical (acoustic) parameters: pitch, amplitude, and duration. While we do not expect prosodic measurements to take the place of other measurements, we show that prosody offers many benefits as an enhancement to phonetic and word-based LID systems by being a semi-independent source of information, resistant to noise corruption, and computationally efficient to implement.

2. METHODS

NST's "discrim" is a prosody-based language discrimination system. As such it was never intended to be a stand-alone solution, but only a component of a complete LID system incorporating more traditional measures of phonetic events and word recognition. The discrim system consists of an acoustic front end which extracts pitch and amplitude information as a function of time, a prosodic analysis unit which performs syllable segmentation and extracts pitch and amplitude contour information on a syllable by syllable basis, a statistical module which computes inter-syllable (timing related) relationships in the pitch and amplitude information, a training module which collects histograms of various features or feature pairs, and a discrimination module which computes log likelihood ratio functions from histograms and uses the log likelihood ratio functions to evaluate "unknown" input in a pairwise discrimination task.

We have examined discrimination between pairs of English, Japanese, Mandarin, and Spanish. Fifty-second speech files from roughly 90 speakers were used for each language. The files were divided into 2 groups of 45 speakers each, designated "TR" and "DF." A modified jack-knifing scheme, termed "cross-runs," was used. Given the TR and DF data for two languages, A and B, the four cross-runs involve the four possible assignments of train and run sets to the two languages in which the run data does not equal the train data. In evaluating a feature or set of features for a language pair, we typically looked at the minimum over the four cross-runs and average over the four cross-runs of the figure of merit derived on a speaker by speaker basis and percent-correct. The FOM of a feature X for a pairwise discrimination of a set of speakers drawn from two languages A and B is the difference in the means of the LLR given each language divided by the sum of the standard deviations of the LLR given each language, i.e.

$$\mathrm{FoM}(X, A, B) = \frac{E(\mathrm{LLR}(X|A)) - E(\mathrm{LLR}(X|B))}{\mathrm{StD}(\mathrm{LLR}(X|A)) + \mathrm{StD}(\mathrm{LLR}(X|B))}$$

Evaluating features according to the minimum, in particular, helps eliminate features that may prove "unstable", i.e. detrimental to system performance when new data with a different style mix is presented.

Based on our own research and research reported in the linguistic and phonetic literature, we implemented a large set of possible prosodic feature measurements designed to capture pitch and amplitude contours on a syllable-by-syllable basis. Considerable effort was devoted to defining feature measurements that were insensitive to overall amplitude, overall pitch, and speaking rate. Individual features were evaluated by deriving a log likelihood ratio function for the given feature and language pair and then evaluating the effectiveness of that function as a discriminator.

The discrim system currently measures 224 individual features of which normally only a small subset are used in the training and discrimination modules. Individual features can be combined into feature pairs. Features are running averages, deltas, standard deviations, and correlations of measures in several classes:

- Pitch Contour (shape of pitch contour on a syllable).
- Differential Pitch (pitch differences – mid point or max – one syllable to the next).
- Size (distance between syllables and syllable duration).
- Differential Size (differenced distance between syllables and syllable duration).
- Amplitude (shape of amplitude contour on a syllable).
- Differential Amplitude (amplitude differences – mid point or max – one syllable to the next).
- Rhythm (low frequency FFT of amplitude envelope; syllables per second within breath group).
- Phrase Location (initial/mid/final in breath group; relative phrase position based on syllable distance ratios).

3. RESULTS

We present results from pairwise LID tasks with English, Spanish, Japanese and Mandarin using the OGI-TS database. These four languages were chosen since they represent the traditional categories of stress-timed, syllable-timed, mora-timed and tone languages. The set of features used in this study consists of 220 features: 47 single features, chosen to span the classes above and to include features previously found to be useful in LID, and 173 paired features, chosen to span the meaningful pairs of classes from the list above and to include previously useful pairs.

For each language pair we represented each feature by the minimum percent correct observed in the four cross-runs. We classified each of the 220 features according to its class or class-pair and found the "best" feature (highest minimum over 4 runs) in each class or class-pair. In the resulting matrices [Tables 1-6], the column and row headings "sg" refer to single features from the respective classes (Rhythm, Phrase Location, differential Size, Size, differential Pitch, Pitch, differential Amplitude, Amplitude). Other entries are for paired features. All classes with features performing at or above 70% are in bold type face.

Most prominent in the charts is the strength of pitch (P and to a lesser extent dP) both alone and in combination with almost every other feature. Overall, we have found combinations of L and dP and of L and P to be the most important in LID. The weakest distinctions involve amplitude and differential amplitude (in italics), suggesting that using amplitude features is a very poor LID strategy in that differences from speaker to speaker could cause discrimination results worse than chance.

The strength of pitch measures may be due to pitch as a signal offering more possibilities for discrimination, including many aspects of absolute and relative pitch, pitch change, slope, and curvature over different portions of the syllable. The strength of pitch measures may also derive from differences in perception of pitch, amplitude and duration. Estimates by Laver [3] suggest that the just-noticeable differences in pitch discrimination offer considerably more distinction possibilities than in duration discrimination. In the typical F0 range, 50-480Hz, the just-noticeable difference is +/– 1Hz. The just-noticeable difference between durations of individual speech-segments (which range between 30 msec and 300 msec) is 10-40 msec. Thus we could argue that there are effectively roughly 430 levels of pitch compared to roughly 25 or fewer usable levels of duration.

The results of some of the strongest individual parameters reflect correlations between perceptually based observations from the linguistic literature and specific physical measurements. Results for English vs. Other are presented in greater detail in [2] but are discussed below as they relate to language-pair results within prosodic parameter classes.

	sg	**Rh**	**L**	**dS**	**S**	**dP**	**P**	**dA**	**A**
sg		58	55	65	55	57	68	*46*	55
Rh	58	60	59	60	58	59	63	60	61
L	55	59	*50*	55	60	62	67	*49*	*49*
dS	65	60	55	63	60	58	**73**	60	*39*
S	55	58	60	60	*50*	56	**72**	55	53
dP	57	59	62	58	56	57	**73**	55	*41*
P	68	63	67	**73**	**72**	**73**	**73**	65	58
dA	*46*	60	*49*	60	55	55	65		*34*
A	55	61	*49*	*39*	53	*41*	58	*34*	*31*

Table 1: English vs. Japanese: Highest Minimum % Correct in Cross-runs

English vs. Japanese: English is a non-tonal stress-timed language, while Japanese is a mora-timed pitch accent language. English tends to be more varied in overall pitch, has different shaped syllable pitch contours, and has different timing features from Japanese. These features are reflected in the strength of measures including pitch peak location, raw syllable pitch change, syllable pitch contour curvature, and delta distance between syllables. That is, P, dS and their combinations.

	sg	Rh	L	dS	S	dP	P	dA	A
sg		59	67	65	64	63	**75**	51	54
Rh	59	64	63	65	69	59	**79**	56	53
L	67	63	56	68	68	67	**79**	54	56
dS	65	65	68	69	64	**72**	**83**	63	54
S	64	69	68	64	66	65	**80**	64	62
dP	63	59	67	**72**	65	59	**71**	*50*	47
P	**75**	**79**	**79**	**83**	**80**	**71**	**76**	**78**	**74**
dA	51	56	54	63	64	*50*	**78**		55
A	54	53	56	54	62	47	**74**	55	*38*

Table 2: English vs. Mandarin: Highest Minimum % Correct in Cross-runs

English vs. Mandarin: Mandarin is a tone language, which is reflected in the strength of P (and dP) parameters against all other languages. The strongest measures in English vs. Mandarin involve pitch slope and delta distance between syllables, the latter reflecting differences in timing between the two languages.

	sg	**Rh**	**L**	**dS**	**S**	**dP**	**P**	**dA**	**A**
sg		63	55	61	54	61	62	*45*	52
Rh	63	65	63	62	62	65	67	59	62
L	55	63	*50*	69	63	64	63	*44*	52
dS	61	62	69	69	62	67	**71**	*50*	51
S	54	62	63	62	55	59	63	55	52
dP	61	65	64	67	59	61	61	52	54
P	62	67	63	**71**	63	61	66	55	54
dA	*45*	59	*44*	*50*	55	52	55		*50*
A	52	62	52	51	52	54	54	*50*	*27*

Table 3: English vs. Spanish: Highest Minimum % Correct in Cross-runs

English vs. Spanish: Spanish is described as a syllable-timed language. Timing differences from English are reflected in strong Rh and dS parameters, including low frequency power spectrum at 6Hz, raw and delta syllable duration, delta distance between syllables, and syllable pitch slope.

Japanese vs. Mandarin: Japanese, a mora-timed pitch accent language, is quite different from Mandarin, a tone language. The overall greater pitch variation of Mandarin is reflected in strong measures of pitch change, particularly in the early part of the syllable. Timing measures are strong only in combination with pitch measures.

	sg	**Rh**	**L**	**dS**	**S**	**dP**	**P**	**dA**	**A**
sg		60	62	59	66	67	**71**	*45*	56
Rh	60	62	65	62	69	67	66	59	59
L	62	65	55	63	64	62	**72**	*45*	56
dS	59	62	63	57	59	68	**77**	54	57
S	66	69	64	59	68	67	**72**	69	65
dP	67	67	62	68	67	66	**72**	*48*	*43*
P	**71**	66	**72**	**77**	**72**	**72**	**74**	**78**	**71**
dA	*45*	59	*45*	54	69	*48*	**78**		*40*
A	56	59	56	57	65	*43*	**71**	*40*	*44*

Table 4: Japanese vs. Mandarin: Highest Minimum % Correct in Cross-runs

	sg	**Rh**	**L**	**dS**	**S**	**dP**	**P**	**dA**	**A**
sg		59	*47*	56	59	61	**71**	*36*	58
Rh	59	58	59	59	58	65	**73**	59	63
L	*47*	59	*46*	56	60	**77**	**72**	*36*	59
dS	56	59	56	53	61	**72**	**76**	*38*	60
S	59	58	60	61	55	67	**72**	55	61
dP	61	65	**77**	**72**	67	58	**70**	66	66
P	**71**	**73**	**72**	**76**	**72**	**70**	**76**	68	69
dA	*36*	59	*36*	*38*	55	66	68		56
A	58	63	59	60	61	66	69	56	*48*

Table 5: Japanese vs. Spanish: Highest Minimum % Correct in Cross-runs

Japanese vs. Spanish: Japanese mora-timing is often considered a type of syllable-timing; Japanese and Spanish are therefore similar in timing and are not easily discriminated based on rhythm and duration measures. Though overall pitch variation is limited in both languages, the difference is observable, as reflected in strong measures of pitch peak location, average pitch slope, and pitch change, particularly in the later part of the syllable.

	sg	**Rh**	**L**	**dS**	**S**	**dP**	**P**	**dA**	**A**
sg		62	65	59	65	65	**80**	*43*	57
Rh	62	69	65	65	68	68	**79**	62	67
L	65	65	54	59	60	67	**86**	*45*	63
dS	59	65	59	62	55	68	**83**	*46*	60
S	65	68	60	55	60	68	**85**	67	66
dP	65	68	67	68	68	60	**77**	*34*	57
P	**80**	**79**	**86**	**83**	**85**	**77**	**78**	77	**74**
dA	*43*	62	*45*	*46*	67	*34*	**77**		54
A	57	67	63	60	66	57	**74**	54	*50*

Table 6: Mandarin vs. Spanish: Highest Minimum % Correct in Cross-runs

Mandarin vs. Spanish: The overall flat pitch of Spanish offers good discrimination against Mandarin, reflected in very strong pitch measures, particularly pitch change over the syllable. Although timing differences are stronger than for Japanese vs. Spanish, the strongest measures involving distance between syllables are strong only in combination with pitch measures.

4. DISCUSSION

Although the most successful LID system should include segmental as well as prosodic cues, our results show that prosodic cues alone can successfully distinguish between some language pairs. The combination of prosodic information from our system with an existing phonetic LID system resulted in improved performance over both original systems for the three language pair English vs. Other [8].

We have found that prosodic measures can be successfully used in distinguishing between different language types, but the discrimination success rate and the performance of particular features are language-pair specific. In the work reported here and in ongoing work with additional languages, we have found that we can to some degree predict the usefulness of a particular feature or feature class for a pair of languages based on their diachronic relationship or language families, synchronic categorization, syllable structure, amount of pitch activity, and strictness of pitch activity.

The results suggest that prosodic parameters are highly useful in automatic LID, but the statistical relationships between a number of individual features deduced from timing and pitch measurements are needed to begin to capture such complex perceptual events as rhythm. Based on a cross-language study of rhythm, Dauer [1] concluded that "The difference between stress-timed and syllable-timed languages has to do with differences in syllable structure, vowel reduction, and the phonetic realization of stress and its influence on the linguistic system." Our findings support this argument: while we have not been able to capture a measure of the complex timing types (that is, stress-timed versus syllable-timed), we have found a number of rhythm-related physical measures that are useful in LID tasks. By considering a very large set of individual prosodic features and combinations of prosodic features, we have captured stronger prosodic cues for LID tasks than previously achieved. The correlation between results from specific features and expectations based on the pitch variation and timing structure of a language suggests that familiarity with the linguistic specifics of a language allows us to make predictions about the usefulness of particular prosodic features in the discrimination of a given language-pair.

Among the current system's strengths are:

- it requires little manual input during training, only the true identity of the language being spoken;
- it is relatively computationally efficient, running faster than real-time on a small SUN workstation;
- the raw features it extracts are relatively immune to noise and other corruption thus rendering the system particularly suitable for noisy or corrupt data and relatively insensitive to changes in transmission channel.

The main technical weaknesses of the current system are that

- the probability distributions are only estimated for single features and feature pairs, not for more complex combinations;
- the simple summation of many log likelihood ratio functions is thought not to be an optimum strategy for decisions based on multiple features.

These issues are currently being addressed.

5. REFERENCES

1. Dauer, R., "Stress-timing and syllable-timing re-analyzed," *Journal of Phonetics* 11: 51-62, 1983.
2. Hutchins, S., and Thymé-Gobbel, A., "The role of prosody in language identification," *Proceedings of the 15th Annual Speech Research Symposium*, Johns Hopkins University, Baltimore, MD, 1995.
3. Laver, J. *Principles of Phonetics*. Cambridge Textbooks in Linguistics, Cambridge University Press, 1994.
4. Muthusamy, Y.K., Barnard, E., and Cole, R.A. "Reviewing automatic language identification," *IEEE Signal Processing Mag.*, vol. 11, no. 4: 33-41, 1994.
5. Muthusamy, Y.K, *Segmental approach to automatic language identification*, Ph.D. thesis, Oregon Graduate Institute of Science & Technology, 1993.
6. Muthusamy, Y.K., Cole, R.A., and Oshika, B.T. "The OGI multi-language telephone speech corpus," *Proceedings International Conference on Spoken Language Processing* 92, Banff, Alberta, Canada, 1992.
7. Savic, M., Acosta, E., and Gupta, S., "An automatic language identification system," *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing 91*, Toronto, Canada, 1991.
8. Zissman, M., and Martin., A. "Language Identification Overview," *Proceedings of the 15th Annual Speech Research Symposium*, Johns Hopkins University, Baltimore, MD, 1995.
9. Zissman, M., "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no.1:31-44, 1996.