LANGUAGE IDENTIFICATION FROM PROSODY WITHOUT EXPLICIT FEATURES

Fred Cummins, Felix Gers, Jürgen Schmidhuber
Istituto Dalle Molle di Studie sull'Intelligenza Artificiale
Corso Elvezia 36
CH 6900 Lugano
Switzerland
{*fred,felix,juergen*} *@idsia.ch*

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

10.21437/Eurospeech.1999-96

ABSTRACT

Most current language identification (LID) systems make little or no use of prosodic information, despite the importance of prosody in LID by humans. The greatest obstacle has been that of finding an appropriate feature set which captures linguistically relevant prosodic information. The only system to attempt LID entirely on the basis of prosodic variables uses a set of over 200 features which are selected and combined in a task-specific manner [12]. We apply a novel recurrent neural network model to the task of pairwise discrimination among languages. Network inputs are limited to delta-F_0 and the first difference of the band limited amplitude envelope. Initial results are based on all pairwise combinations of English, German, Japanese, Mandarin and Spanish, with 90 speakers per language.

Keywords: Language identification, Recurrent neural networks, prosody

1. PROSODY AND LANGUAGE IDENTIFICATION

Most current approaches to automatic language identification use some form of segment recognition, and base subsequent identification on segmental and phonotactic probabilities [7]. A few models incorporate a limited amount of prosodic information [4], but in the absence of a well-defined set of prosodic features, prosody is still of very limited utility. This is despite evidence that prosody serves an important role in human identification of languages, both for adults [10] and even more so in infants [1].

Although the relationship between acoustic patterns and underlying segments or features is complex, much successful work has been done on establishing the nature of this relationship. The underlying segmental or featural inventories are usually reasonably uncontroversial. The same cannot be said for prosodic units, which remain elusive even in well-studied languages like English and Japanese [9]. Likewise, the relationship between concrete or physical prosodic variables, such as F_0 modulation or amplitude envelope variation, and underlying linguistic units is poorly understood and further complicated by the influence of a host of extra-linguistic co-determinants, such as emotional state, gender, speaking style, etc.

In the largest study which evaluated the potential of exclusively prosodic features in language identification, Thymé-Gobbel and Hutchins [12] used 47 single features and a further 173 feature pairs in pairwise discrimination tasks among the languages English, Spanish, Japanese and Mandarin. Features included averages, deltas, standard deviations and correlations of measures based on pitch, syllable duration, amplitude, low frequency FFT of the amplitude envelope and phrase location within a breath group. For each feature or feature pair, a likelihood statistic based on histograms was computed and used for discrimination on test data. Overall, they found pitch-based features to be of most utility, either alone, or in combination with other features. Amplitude-based features fared worst, often leading to performance below the level of chance. Although best performance was very encouraging, the approach suffered from the serious drawback of relying on combinations from a very large set of features, with different features proving maximally effective in different pairwise discrimination tasks.

A major problem with this approach is the task of deciding among the multitude of candidate features. When combinations of features are to be used, comparative evaluation of features sets becomes computationally intractable. In the present study, we seek to establish whether discrimination among languages based on prosodic variables is possible without any *a priori* decisions about the form of underlying features. To this end, we

train a novel recurrent network to discriminate among languages, providing *only* either the first difference of F_0 or the first difference of the bandpass filtered amplitude envelope as input, and the language identity as a training signal.

2. EXPERIMENTS

2.1 Speech processing

We use a subset of the OGI Multi-Language Telephone Speech Corpus [8], restricting our attention to the five languages English, Japanese, Spanish, Mandarin Chinese and German. The first four languages were those used in [12], and were selected because they include stress-, syllable- and mora-timed languages, and languages with lexical tones, stress accents, and pitch accents.

The OGI corpus provides recordings acquired over commercial telephone lines of speakers' responses to an automatically generated series of prompts. The prompts elicit lexically constrained responses (e.g. "Please recite the seven days of the week"), topic-specific but otherwise unconstrained responses (e.g. "Tell us something that you like about your home town") as well as about 60 seconds of completely unconstrained speech per caller. For each network simulation, 50 speakers of each language were randomly assigned to the training set, and 20 speakers each to the validation and test sets. All topic-specific utterances and short unconstrained utterances from each speaker were used in training, validation and testing of individual networks. Generalization was then further tested by presenting long (max 50 sec) files of unconstrained speech to a committee of 10 networks and taking the average network output over the final 0.5 sec.

Log F_0 was estimated for each 1 ms of speech. This estimate was first differenced, downsampled to 100 Hz using a sliding rectangular window, smoothed using a simple 15 point rectangular window average, and rescaled to lie within the range [-1,1]. We refer to this input as ΔF_0.

The amplitude envelope was computed by filtering the speech with a low-order Butterworth bandpass filter centered at 1000 Hz with a bandwidth of 500 Hz. Results of both Cummins (1997), and Scott (1993), suggest that the amplitude variation in this frequency range is important for the perceived rhythm

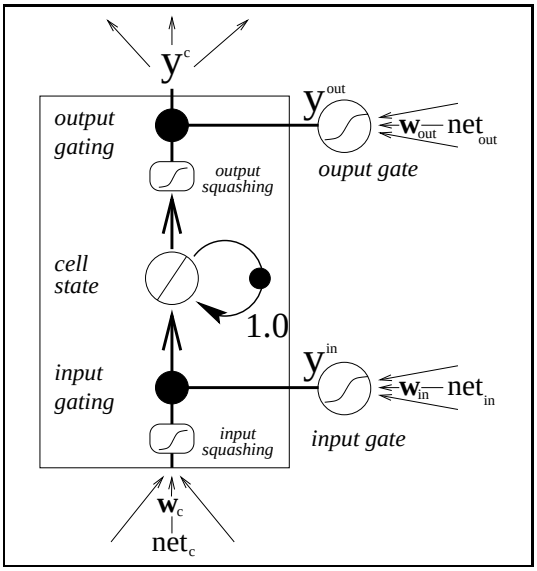

Figure 1: One LSTM block containing a single cell.

of speech. The filtered speech was rectified by taking absolute values of each sample and then smoothed using a Butterworth lowpass filter with cut-off at 10 Hz. It was then first-differenced, downsampled to 100 Hz, resmoothed and rescaled in the same manner and using the same parameters as for ΔF_0. This input will be called ΔEnv. Network inputs thus consist of a time series of either ΔF_0 or ΔEnv at 100 Hz.

2.2 Network training

Recurrent neural networks which automatically extract informationally relevant features from on-line input offer the promise, in principle, of finessing the problem of choosing effective inputs from a large set of candidate featural representations. Conventional recurrent neural networks, however, suffer from the severe limitation that information which is distributed over more than about 10–12 time steps (0.1 sec for the present data) cannot, in general, be effectively used [5]. An attempt to overcome these limitations is the Long Short-Term Memory (LSTM) model, first presented in Hochreiter and Schmidhuber (1997). In an LSTM network (Fig 1), conventional hidden units are replaced by memory *blocks* containing one or more memory *cells*. The heart of a cell is a linear unit with a fixed self-recurrent connection with weight of 1.0, which ensures that activation in the cell remains there in the absence of any other input. Activation flowing into the unit is gated by an *input gate*, which is

a sigmoidal unit with activation ranging over [0,1]. The net input to each cell in a block is multiplied by the activation of the input gate, allowing the input gate to decide what information the cell is exposed to. After gating, cell input is squashed by a centered sigmoid. Likewise, cell output is first squashed by a centered sigmoid, and then gated by the activation of an *output gate*. When activation flows forward in the network, the input gate decides which information should be allowed into the cells, while the output gate decides when the cell should contribute to the net input of other units.

Network training is a combination of truncated Back Propagation Through Time and Real-Time Recurrent Learning. Details are provided in [6]. The constant activity, and hence constant error, stored in the cells allows an LSTM network to retain information over indefinitely long periods of time. We used networks having three blocks, each with a single input and output gate and two cells. Full details are given in [3].

During training, networks were presented at each time step with a target (1 or 0) indicating the language being presented. A sequence was judged to have been classified correctly if the average output for the last 50 inputs (0.5 sec) was on the correct side of 0.5. After each training epoch, weights were frozen and performance on the validation set measured. Weights were stored when performance on the validation set was optimal (max 60 epochs), and performance was tested on the independent test set. Finally, committees of 10 networks were presented with longer files of unconstrained speech and network outputs were averaged. Averaging network outputs has the effect of favoring more confident networks which exhibit more extreme outputs.

3. RESULTS

	Ger	Spa	Jap	Man
Eng	52 (4)	56 (5)	50 (5)	59 (5)
Ger	-	51 (3)	55 (4)	58 (3)
Spa	-	-	59 (4)	50 (3)
Jap	-	-	-	63 (3)

Table 1: Mean percent correct (s.d.) for each pairwise discrimination task, given only ΔEnv as input.

	Ger	Spa	Jap	Man
Eng	52	62 [52]	57 [55]	58 [54]
Ger	-	51	58	65
Spa	-	-	66 [58]	47 [57]
Jap	-	-	-	60 [56]

Table 2: Mean percent correct for a committee of 10 trained networks on long samples of unconstrained speech, given only ΔEnv as input. Best comparable results from [12] are given in brackets.

Table 1 gives the mean percent correct for individual networks (n=10) trained on each pairwise comparison using ΔEnv only as input, and tested on 10 second files of speech from novel speakers. Table 2 gives performance of a committee of 10 such networks presented with long sound files of unconstrained speech. The most comparable figures from [12] are also provided in brackets. These are the best results obtained using their set of amplitude-based features alone. Overall performance using this input variable is modest. The committee results obtained using LSTM are, in general, somewhat better than those obtained in [12] using explicit features.

	Ger	Spa	Jap	Man
Eng	56 (4)	50 (2)	63 (5)	63 (3)
Ger	-	54 (3)	69 (3)	69 (4)
Spa	-	-	60 (3)	62 (4)
Jap	-	-	-	50 (2)

Table 3: Mean percent (s.d.) correct for each pairwise discrimination task, given only ΔF_0 as input.

	Ger	Spa	Jap	Man
Eng	55	52 [62]	62 [68]	62 [75]
Ger	-	54	72	70
Spa	-	-	71 [71]	63 [80]
Jap	-	-	-	44 [71]

Table 4: Mean percent correct for a committee of 10 trained networks on long samples of unconstrained speech, given only ΔF_0 as input. Best comparable results from [12] are given in brackets.

Tables 3 and 4 present similar results using ΔF_0 only as input. The best results from [12] are those obtained using pitch-based features alone. Performance as a whole is rather bet-

ter than that obtained using ΔEnv. Mandarin and Japanese, in particular, are well differentiated from the Indo-European languages, though not from one another. In this case, the explicit features of [12] do a somewhat better job than LSTM.

4. DISCUSSION

These results are consistent with the literature which finds that prosodic variables can contribute, though modestly, to automatic language discrimination. As in previous studies, we find F_0 to be a more effective discriminant variable than amplitude envelope modulation. However, the present results suggest that envelope modulation may be more effectively exploited than heretofore.

The LSTM network employed here has so far been tested only on symbolic and artificial data. Its performance on these difficult real-world data suggests that recurrent networks may indeed have a role to play in the automatic identification of unknown features which are germane to a given discrimination task. An immediate research goal is the development of means for discovering the characteristics of the input which are exploited by the trained networks. This, in turn, may contribute to linguistic efforts to identify language-specific prosodic features.

Much remains to be done. Neither extant systems, nor the present results, reflect the putative importance of the role played by prosody in language identification by humans.

5. ACKNOWLEDGEMENTS

This work was funded by SNF grant 21-49144.96 (Long Short-Term Memory).

6. REFERENCES

[1] Josiane Bertoncini, Caroline Floccia, Thierry Nazzi, and Jacques Mehler. Morae and syllables: rhythmical basis of speech representation in neonates. *Language and Speech*, 38(4):311–329, 1995.

[2] Fred Cummins. *Rhythmic Coordination in English Speech: An Experimental Study.* PhD thesis, Indiana University, Bloomington, IN, 1997. Also Technical Report 198, Indiana University Cognitive Science Program.

[3] Fred Cummins, Felix Gers, and Jürgen Schmidhuber. Automatic discrimination among languages based on prosody alone. Technical Report IDSIA-03-99, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Lugano, CH, 1999.

[4] Timothy J. Hazen and Victor W. Zue. Segment-based automatic language identification. *Journal of the Acoustical Society of America*, 101(4):2323–2331, 1997.

[5] J. Hochreiter. Untersuchungen to dynamischen neuronalen Netzen. Diploma thesis, Technische Universität München, 1991.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] Yeshwant K. Muthusamy, Etienne Barnard, and Ronald A. Cole. Automatic language identification: A review/tutorial. *IEEE Signal Processing News*, October 1994.

[8] Yeshwant K. Muthusamy, Ronald A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *Proceedings of the International Conference on Spoken Language Processing 1992*, Banff, Canada, 1992.

[9] Janet B. Pierrehumbert and Mary E. Beckman. *Japanese Tone Structure.* Linguistic Inquiry Monographs. MIT Press, Cambridge, Ma, 1988.

[10] Franck Ramus and Jacques Mehler. Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, 105(1):512–521, 1999.

[11] Sophie K. Scott. *P-centers in Speech: An Acoustic Analysis.* PhD thesis, University College London, 1993.

[12] Ann Thymé-Gobbel and S. E. Hutchins. On using prosodic cues in automatic language identification. In *Proceedings of the 1996 International Conference on Spoken Language Processing*, volume 3, pages 1768–1771, Philadelphia, PA, 1996.