# AUTOMATIC LANGUAGE IDENTIFICATION: USING INTONATION AS A DISCRIMINATING FEATURE

V.F. Leavers, K. Wiehler, C.E. Burley
Electrical Engineering Division, Manchester University, Dover Street, Manchester, M13 9PL, England

April 23, 1996

## Abstract

Current research into automatic language identification systems sees the problem as being related to speaker independent speech recognition and speaker identification. In particular, speaker indentification methods appear to outperform all other other methods and the incorporation of prosodic information has contributed only marginally to their success. This is a counterintuitive result suggesting that perhaps the brute-force application of standard available pattern recognition methods is inappropriate, not least because it ignores the linguistic cues that human beings use so easily and efficiently. It has been proposed that an attempt to rank parameter extraction with respect to a taxonomy of linguistic complexity would give results more in keeping with our own abilities to discriminate between various languages. For example, the pressure of discrimination concerning grossly different languages such as Mandarin Chinese and English would be low compared to that associated with an attempt to distinguish between two quite similar languages such as Dutch and German. The present work aims to differentiate between the two broadest groups separating tone and stress languages by using parameters which best model the linguistic differences between those groups. In particular, the supra-segmental feature of intonation is modelled as a memory effect which can be measured using the Hurst exponent.

## 1. INTRODUCTION

Automatic language identification is the problem of identifying the language being spoken from a sample of speech by an unknown speaker. The rapid growth in international trade and communications brings with it an ever increasing need for such services. Of all the possibilities for automatically replicating human skills, automatic language identification is perhaps unique in its potential to outperform the human operator. This is because human performance and competence vary widely across social, national, cultural and educational classes [1]. However, irrespective of their overall competence, humans can determine, within seconds of hearing speech, whether that speech belongs to a language that they know. If it does not they can often make subjective judgements concerning its similarity to a language they do know. In order to do this they appear to be exploiting the fact that different languages have distinctive supra-segmental prosodic signatures.

Automatic language identification is clearly an area of research which requires much interdisciplinary collaboration. However, as evidenced by the most recent review [2], the major part of the research is carried out by the speech recognition/signal processing community where automatic language identification is seen as being related to speaker independent speech recognition and speaker identification. Conventionally, the extraction of acoustic features coupled with the use of various standard pattern recognition methods form the basis of the many proposed algorithms. For example, current trends centre on the use of Vector Quantisation [3], Hidden Markov Models and Neural Networks [4]. While considerable progress has been made, many problems remain and the recognition accuracy that is generally achievable falls short of that required by many practical applications. In addition, the need for lengthy speech samples [2] and the high computational complexity and long run

times of the algorithms precludes their use in real world applications.

It has been proposed that an attempt to rank parameter extraction with respect to a taxonomy of linguistic complexity would give results more in keeping with our own abilities to discriminate between various languages [5]. For example, the pressure of discrimination concerning grossly different languages such as Mandarin Chinese and English would be low compared to that associated with an attempt to distinguish between two quite similar languages such as Dutch and German. The use of such a taxonomy would also mean that systems could be efficiently tailored to the task in hand offering potential reductions in the computational complexity of the system as a whole.

## 2. DATA BASE

English and Chinese were chosen as being typical examples of tone and stress driven languages, respectively. $F_0$ patterns can be either linguistic or emotional in origin irrespective of the language being spoken. Thus, in order to concentrate on the linguistic differences between the two languages, subjects were asked to read passages with little or no emotional content.

The database for the analysis consisted of speech samples taken from 10 different male speakers; 5 Chinese and 5 English. The data were recorded on an audio cassette and then sampled with 16bit resolution and 32kHz sampling frequency. Finally, the speech samples were down-sampled by 4:1 to achieve 20 samples with a resolution of 16bit, a sampling frequency of 8kHz, and a duration of 50 seconds. For a more detailed account see [8].

## 3. CHOICE OF PARAMETERS

English uses stress at the word level where primary stress is indicated by an increase in the intensity of the fundamental frequency, $F_0$, and in the duration of the syllable. However, while intensity modulation is a distinguishing factor, in order to be a useful discriminator, the intensity measurements would need to be made in a special noise free environment. This is not possible in real world situations. It was therefore decided not to use changes in the intensity patterns as a discriminating parameter. At the level of the sentence, the $F_0$ patterns of English are dominated by intonation; this varies according to whether a sentence is declarative, interrogative, exclamatory etc. [6]. Hence, parameters that measure the use of intonation and the variation in the duration of voiced segments were chosen as discriminatory factors.

At the word level in Chinese, the use of changes in $F_0$, or tone (as opposed to intonation), is very different. Tones assigned to syllables distinguish lexical items. At the level of the sentence, coarticulation causes adjacent tones to influence each other. There is a rule goverened system of modification of the underlying tone contour when it occurs in a particular tonal environment [7]. Although these changes are dictated at the level of the sentence, they are place dependent and do not affect the complete tonal signature of the sentence in the same way that intonation modulates the $F_0$ pattern of a sentence in English. The rate of change of $F_0$ was chosen as being the parameter most sensitive to the characteristic tone changes of Chinese.

While it is relatively simple to extract and interpret the parameters which model the rate of change of $F_0$ and the duration of the voiced segments, it is more difficult to model a feature such as intonation. This is because it is a global feature, and it is necessary to consider the production of a whole phrase or sentence. The production of a sound within a sentence at a given point in time is the result of a stream of interconnected events. The position in time of those events is significant. Thus, it is important to realise that each observation is not independent; it carries with it a "memory" of all preceding associated events. The length of the memory effect will be dictated by the type of $F_0$ contour being produced. In order to begin and execute a sentence in which the overall pitch contour is modulated by the effect of intonation, the speaker must plan ahead to the end of the sentence. On the other hand, in order to use tone, the speaker need only deal with the local effects of the coarticulation of adjacent tones. Hence, for a tone language, such as Chinese, the memory effect is short term. For a stress language, such as English, the memory effect is long term. These effects can be measured by using the Hurst exponent [9] to quantify the degree of persistance or anti-persistance of the rate of change of fundamental frequency.

## 4. RESULTS

The results of the experiments are shown in graphical form in Figures 1, 2 and 3. Fig. 1. shows the two plots of the normal distributions fitted to the observed data for, on the left, rate of change of $F_0$ and, on the right, average voiced segment duration. The vertical bars are of two standard deviations width. The area under the curve between these lines is 68% of the total distribution. As can be seen neither parameter alone gives sufficiently robust discriminating power. Fig. 2 shows the surface plot of the rate of change of $F_0$ against the average duration of the voiced segments. As can be seen, an increase in discriminating power has been achieved. Fig. 3 shows various plots of the results used to calculate the Hurst exponent. For a persistent or anti-persistent series, the roughly linear portion of the plot indicates there is a memory effect. The slope of this linear portion is the Hurst exponent. When the linearity of the plot breaks down this is because the memory effect has been lost. A value of 0.5 for the Hurst exponent indicates randomly varying data that do not exhibit long term trends. A persistent series is indicated by a value of the Hurst exponent greater than 0.5; an anti-persistent series is indicated by a value less than 0.5. As can be seen in Fig. 3, the values of the Hurst exponent for English are above 0.5. Those of Chinese are below 0.5.

## 5. DISCUSSION AND CONCLUSIONS

In summary, the work supports the view that the exploitation of an appropriate system of linguistic taxonomy may be the way forward in developing efficient and generic automatic language identification systems. To this end, the work has demonstrated that it is possible, without the use of a computationally complex pattern classifier, to extract a set of sufficiently robust discriminating parameters in order to discriminate between English and Chinese. In addition the calculation of each parameter is independent of the calculation of either of the other two and may be implemented in parallel. This is an important consideration if run times are to be commensurate with real world applications.

## References


[1] Muthusamy Y.K., Cole R.A. and Oshika B.T. *Perceptual benchmarks for automatic language identification,* IEEE International Conference on Acoustics, speech and signal processing 94, Adelaide, Australia, April 1994.

[2] Muthusamy Y.K., Barnard E. and Cole R.A., *Reviewing Automatic Language Identification,* IEEE Signal Processing Magazine, October 1994.

[3] Gray R., *Vector Quantisation,* IEEE ASSP 1. pp 4-29 (1984).

[4] Paliwal K.K., *Neural Net Classifiers for Robust Speech Recognition Under Noisy Environments,* IEEE ICASSP 1990.

[5] V.F. Leavers and A.F. Erwood, *The extraction and identification of language specific information-bearing parameters from digitised speech samples,* 3rd UK/Australian International Symposium on DSP for communication Systems, Warick, December 1994.

[6] Pike K., *The intonation of American English,* Ann Arbor, Mich, 1945.

[7] S.J. Eady, *Differences in the $F_0$ patterns of speech: tone language versus stress language,* Language and Speech, Vol 25(1), 1982.

[8] V.F. Leavers, K. Wiehler and C.E. Burley, *Technical report No. 173,* Electrical Engineering Division, Manchester University, September 1995.

[9] Mandelbrot B.B. and Wallace J.R., *Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence,* Water Resources Research, Vol 5(5), October 1969.

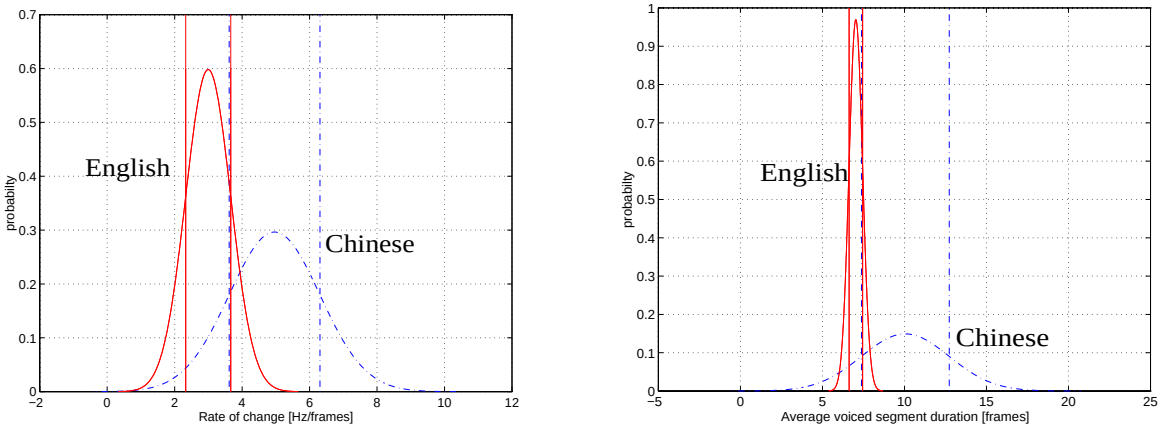


Figure 1

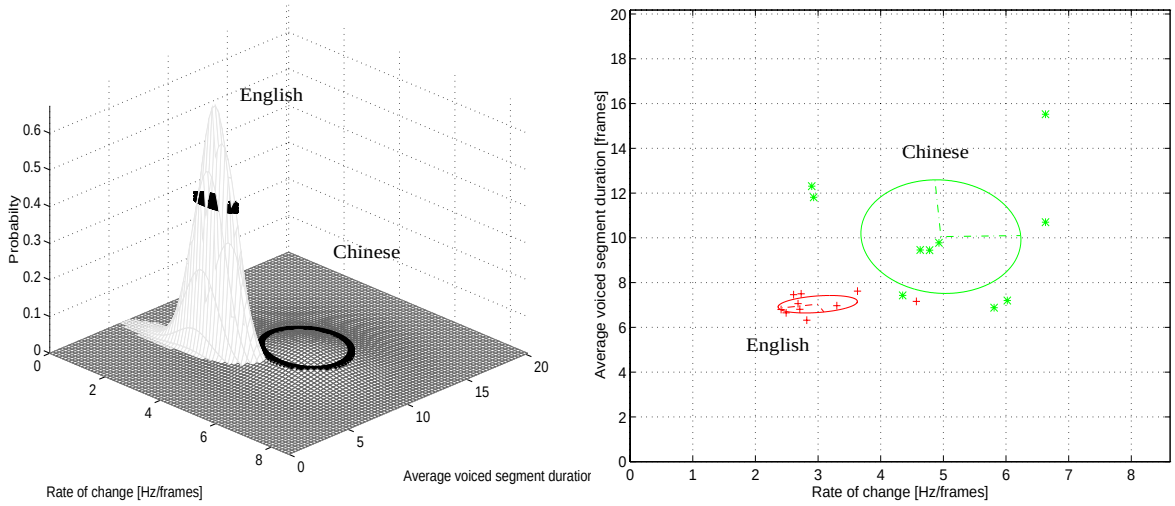


Figure 2

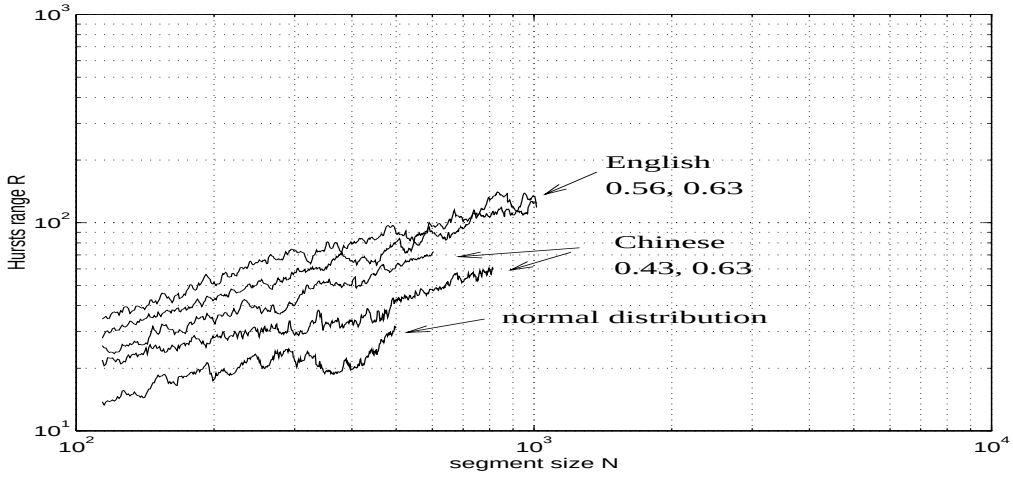


Figure 3