Proceedings of The Institute of Acoustics 2A11

THE AUTOMATIC RECOGNITION OF SPOKEN LANGUAGES USING STATISTICAL TECHNIQUES

A. R. HALL

ELECTRICAL ENGINEERING DEPARTMENT,
UNIVERSITY COLLEGE LONDON

Introduction

This subject concerns the problem of automatically determining, for a given stretch of speech signal, which language is being spoken. The method that has been studied uses samples of conversation to determine the parameters of a stochastic model for each language (the supervised training process). These classification features in the form of transition probabilities are estimated from 1) a sequence of piecewise linear approximations to the fundamental period contour, and 2) a sequence of broad linguistic categories (here termed sound classes) derived from the speech waveform using linear discriminant analysis. To identify the language of an unknown speech sample, a statistical technique is used based on summing the log-probabilities derived from each model.

Language Modelling and Initial Processing

A variety of models has been previously proposed. Transcribed text has provided input samples as starting points in several cases, with criteria derived from words (Menzerath: 1), vowel and consonant patterns (Newman:2, Rau: 3), or from sequences of broad linguistic categories (House and Neuberg: 4). In contrast it has also been shown possible (Atkinson: 5, Ohala and Gilbert: 6) for subjects in listening tests to discriminate between languages on the basis of intonation and durational cues only.

Consequently, it was considered advisable to use some characteristics of both segmental and prosodic information in studying automatic language recognition. In order not to reduce or stylize the prosodic content, only conversational material was used. Recordings of approximately eighty speakers in four languages were obtained (mainly in London language schools) by setting each pair of participants a pictorial quiz to solve by conversation over a telephone link. These four languages were French, Japanese, Persian and Spanish.

Each of the eighty 15-minute samples of speech was converted to a digital binary representation by low-pass filtering at 3500 Hz, sampling at 8 kHz and storing as a sequence of 12-bit words on digital magnetic tape.

The speech was then initially analysed by computer with 32ms time windows taken at 12.5ms intervals. A standard speech production model was assumed (as detailed in Flanagan: 7, for instance). The excitation parameters were derived using cepstrum techniques and Noll's pitch extraction algorithm (8). The vocal tract response, initially in the form of the detailed short-time spectrum, was described by ten spectrum samples logarithmically spaced across the frequency axis. Thus at this stage the speech was in a representation similar to that used in a coarse channel vocoder.

In order to derive features associated with segmental aspects of the speech,

the above spectrum and fundamental period estimates were used to classify each windowed speech section into one of six broad linguistic categories. These 'sound classes' were taken to be silence (including pauses and stop-gaps), vowel sounds, voiced fricatives, other voiced sounds, unvoiced fricatives, and plosive sounds. A standard pattern recognition approach was used - that of linear discriminant analysis (Cooley and Lohnes: 9) - with classification taking place in the resulting discriminant subspace by Euclidean centroid distance. Manually pre-labelled speech provided training and testing samples for this sound class recognition stage. Full details of this and the other techniques involved are given in (Hall: 10). The sound class recognition accuracy was estimated as 61% for the six sound classes, poor by speech recognition standards employing independent semantic information, for instance, but providing a practical approximate transformation into broad linguistic categories when the language is unknown.

The result of the above processing was, for the conversation from each of the eighty speakers, a time sequence of sound classes and associated fundamental period.

The six-state sound class sequence was modified to associate long silences with a seventh state. The final sound class model for each language then comprised a 7x7 matrix of counts of transitions from one state to the next, derived by summing the transition counts from half the speakers in each language. The other speakers were later used to test the language recognition accuracy.

The fundamental period contour was smoothed using a 5-long median smoothing algorithm (Rabiner, Sambur and Schmidt: 11). The contour was then described by a piecewise linear approximation with the line-segment end-points restricted to lie on the contour (Mead: 12). It was the sequence of slopes of these line segments, heavily quantized with up to nine possible values only, that was used as the sequence of states for language modelling based on the period contour. Again a matrix of transition counts provided the model's parameters.

Language Recognition and Conclusions

For large enough speech samples, the conditional transition probabilities may be estimated from the state transition counts. Speech in an unknown language may then be given a measure of similarity to each language model. This is derived by summing the log-probabilities (obtained from the model) of the corresponding sound class or contour slope sequence from the unknown speech. In the testing situation the unknown speech is assigned to that language for which the sum of the resulting sound class and contour slope similarity measures is a maximum.

The language recognition accuracy was estimated using those speech samples which were not used in the model's training stage. A significant accuracy of 64% for four languages was noted. Tests using the sound class and contour slope models separately showed that each model was approximately making an equal contribution to this overall result.

Finally, an experiment was performed to assess the amount of testing data

required to give stable language recognition. The similarity measures were computed at regular time intervals and could be plotted as in fig. 1. It was apparent that the behaviour stabilized after approximately 1 minute of conversation, representing roughly 30 seconds of continuous speech.

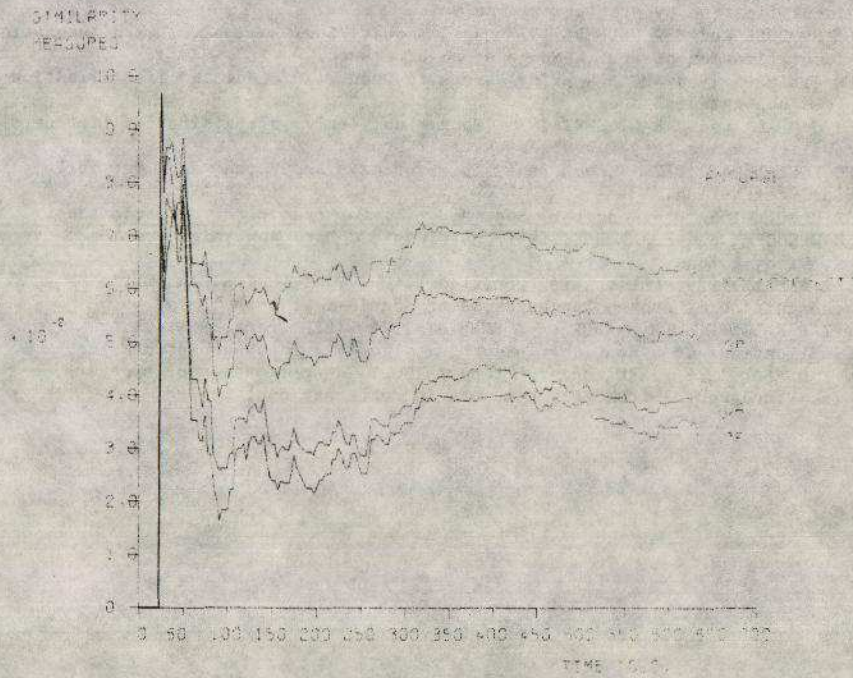

Fig. 1 Plot of Sequential Similarity Measures versus Time

These investigations have demonstrated that both the segmental and prosodic natures of speech separately contain information relevant to recognition of the language being spoken. However, the language recognition methods that have been described are not yet accurate enough to be applied usefully. A major factor is probably the poor sound class recognition. A more accurate derivation of coarse segmental information should improve the language recognition accuracy to enable use in applications such as international conferences.

References

1. P MENZERATH 1950 Jnl. Acoust. Soc. Am., 22. Typology of languages.
2. E B NEWMAN 1951 Am. Jnl. Psychol., 64. The pattern of vowels and consonants in various languages.
3. M D RAU 1974 MA Thesis, Naval Postgrad. Sch., Monterey. Language identification by statistical analysis.
4. A S HOUSE and E P NEUBERG 1977 Jnl. Acoust. Soc. Am., 62. Toward automatic identification of the language of an utterance.
5. K ATKINSON 1968 UCLA Working Papers in Phon., 10. Language identification from nonsegmental cues.
6. J J OHALA and J B GILBERT 1978 Phon. Lab., Univ. of California. Listeners' ability to identify languages by their prosody.
7. J L FLANAGAN 1972 Speech Analysis, Synthesis and Perception. New York: Springer-Verlag.
8. A M NOLL 1967 Jnl. Acoust. Soc. Am., 41. Cepstrum pitch determination.
9. W W COOLEY and P R LOHNES 1971 Multivariate Data Analysis. New York: J. Wiley & Sons.
10. A R HALL 1979 PhD thesis submitted to Univ. of London. Automatic recognition of spoken languages using statistical techniques.
11. L R RABINER, M R SAMBUR, C E SCHMIDT 1975 IEEE Trans. ASSP-23. Applications of a nonlinear smoothing algorithm to speech processing.
12. K O MEAD 1974 JSRU research report 1002. Identification of speakers from fundamental-frequency contours in conversational speech.