Automatic Sound Identification based on Prosodic Listening

Eduardo Reck Miranda

Sony Computer Science Laboratory Paris, 6 rue Amyot, 75005 Paris, France

This paper presents a system for the identification of vocal and vocal-like sounds by means of a classificatory scheme based upon their prosodic attributes.[1] The system firstly segments examples of the sound classes in question and performs a number of analyses on these segments in order to determine their prosodic attributes. Then, it builds a classificatory scheme, which will be used to identify an unknown sound. For example, consider the identification of the language of utterances spoken in English, French and Japanese. In this case, the system establishes a classificatory scheme from samples of speech in these three languages. Then, the system infers the language of unknown utterances by matching their prosodic information with the classificatory scheme. The system is not, however, limited to language identification tasks; it also can identify singing styles and a number of animal calls. Given the appropriate training data, the system is able to identify such sounds with great accuracy.

INTRODUCTION

We are interested in building listening devices for the characterisation and/or identification of spoken utterances, singing, and other sound sequences that resemble the human voice, including a variety of animal calls and synthesised sounds. Previous attempts at devising such listening systems processed a symbolic representation of acoustic phenomena, as opposed to processing the actual acoustic signal; e.g., written text to represent speech and standard musical notation to represent music. A few trials with acoustic signals have been reported in the field of automatic language identification (ALI) systems. Unfortunately these systems often depend upon embedded linguistic knowledge programmed manually [1, 2]. This imposes severe restrictions on such systems, preventing their usability in domains other than ALI, such as automatic identification of singing styles, identification of the mood of the speaker, and sound-based surveillance, monitoring and diagnosis, to cite but a few.

We have devised an experimental system that is able to establish its classificatory scheme autonomously, using language-independent prosodic information extracted automatically from examples of the sound classes in question. The baseline architecture of the system is shown in Figure 1. It is composed of two modules: the Training Module and the Identification Module.

THE TRAINING MODULE

The input to the Training Module is audio samples and the output is a discriminant structure. The audio samples are labelled according to the classes that the system will be required to identify; e.g., the label English for a sample spoken in English , or French for a sample spoken in French. This module is composed of three main procedures: Segmentation, Prosodic Analysis and Assortment. Whereas the task of the *Segmentation procedure* is to chop the input signal into n smaller segments S, the task of the *Prosodic Analysis procedure* is to extract prosodic information from these segments [3]. It performs three types of analysis on each segment S_n, namely Pitch Analysis, Intensity Analysis and Formant Analysis, in order to extract the following information: (i) the standard deviation of the pitch contour of the segment: $\Delta p(S_n)$; (ii) the energy of the segment: $E(S_n)$; (iii) the mean centre frequencies of the first, second and third formants of the segment: $MF_1(S_n)$, $MF_2(S_n)$ and $MF_3(S_n)$; (iv) the standard deviation of the first, second and third formant centre frequencies of the segment: $\Delta F_1(S_n)$, $\Delta F_2(S_n)$ and $\Delta F_3(S_n)$.

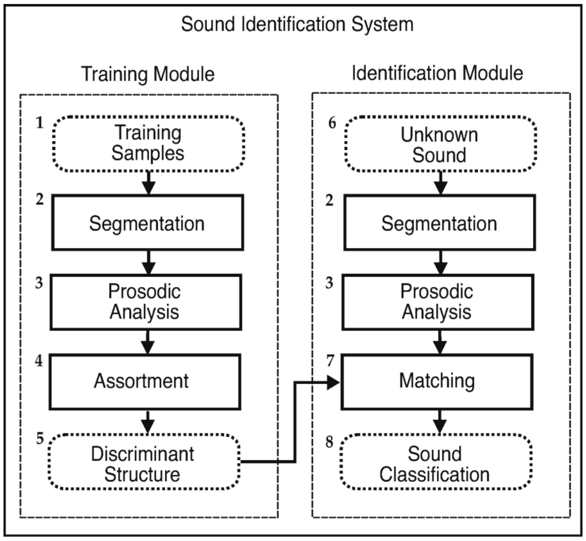

FIGURE 1. The baseline system.

[1] European Patent No. 01400821.3-2218.

The task of the *Assortment procedure* is to build a classificatory scheme (i.e., a discriminant structure), using the extracted prosodic information. The objective here is to identify which prosodic attributes contribute most to differentiate between the given classes, or groups, and this is achieved by means of discriminant analysis [4]. Discriminant analysis is used to build a predictive model of group membership based on observed characteristics, or attributes, of each case. Firstly, the system takes the outcome of the Prosodic Analysis procedure and builds a matrix; each line corresponds to one segment S_n and the columns correspond to the prosodic attribute values of the respective segment: $\Delta p(S_n)$, $E(S_n)$, $MF_1(S_n)$, $MF_2(S_n)$ and so forth. Then, it standardises the values of the columns of the matrix, in order to ensure that scale differences between the values are equalised. Finally, the matrix is submitted to discriminat analysis. Briefly, discriminant analysis works by combining attribute values $Z(i)$ in such as way that the differences between the classes are maximised. The goal of discriminant analysis is to weigh the attribute values in some way so that single new composite attributes, or discriminant scores are generated. These constitute new axes in the space whereby the overlap between the different classes is minimised, by maximising the ratio of the between-class variances to the within-class variances [5]. The weighting coefficients are given by two matrices, namely, the transformation matrix E and the feature reduction matrix f which transforms $Z(i)$ into discriminant vectors $y(i) = f.E.Z(i)$. The result is a discriminant structure of a multivariate data set with several groups. This discriminant structure consists of a number of orthogonal directions is space, along which maximum separability of the groups can occur.

THE IDENTIFICATION MODULE

The task of the Identification Module is to classify an unknown sound based upon a given discriminant structure. The inputs to the Identification Module are therefore the unknown sound to be identified plus the discriminant structure generated by the Training Module. The unknown sound is submitted to the same Segmentation and Prosodic Analysis procedures as for the Training Module, followed by a *Matching procedure*. The task of the Matching procedure is to identify the unknown sound, given its prosodic coefficients and a discriminant structure. Firstly, a matrix is generated from the prosodic coefficients; this matrix is identical to the one generated by the Training Module. Next, the columns of the matrix are standardised. Then the system generates a classification table containing the probabilities of group membership of the elements of this matrix against the given discriminant structure. The probabilities of group membership p_j for a vector x are defined as:

$$p(j \mid x) = \frac{e^{-d_j^2(x)/2}}{\sum_{k=1}^{K} e^{-d_k^2(x)/2}} \tag{1}$$

where K is the number of classes and d_i^2 is a square distance function:

$$d_i^2 = ((x-\mu_i)' \, \Sigma^{-1}(x-\mu_i)) - \log\left(\frac{n_i}{\sum_{k=1}^{K} n_k}\right) \tag{2}$$

where Σ stands for the pooled covariance matrix (it is assumed that all group covariance matrices are pooled), μ_i is the mean for group i and n_i is the number of training vectors in each group. Finally, the classification table is submitted to a confusion procedure which in turn returns the classification of the sound. This procedure generates a confusion matrix, with stimuli as row indices and responses at column indices, whereby the entry at position [*i*][*j*] represents the number of times that response *j* was given to the stimulus *i*.

CONCLUSION

We have trained the system with samples from radio news readings in English, French and Japanese. Then we made queries for samples other than those used for training, but with speakers of the same gender and identical voice timbre. The system was able to identify the language for these queries with 90% accuracy. The system was also successfully trained to identify Gregorian, Tibetan and Vietnamese singing with the same degree of accuracy. We are currently testing the system with a variety of languages, speech styles, singing and other sounds.

REFERENCES

1. Zissman, M.A. and Singer, E. Automatic Language Identification of Telephone Speech Message Using Phoneme Recognition and N-Gram Modeling , in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing - ICASSP 94*, Adelaide, Australia, 1994.
2. Hieronymous, J. and Kadambe, S., "Spoken Language Identification Using Large Vocabulary Speech Recognition", in *Proc. of the 1996 Int. Conf. on Spoken Language Processing - ICSLP 96*, Philadelphia, USA, 1996.
3. Kompe, R., *Prosody in Speech Understanding Systems*, Lecture Notes in Artificial Intelligence 1307, Springer Verlag, Berlin, 1997.
4. Klecka, W.R., *Discriminant Analysis*, Sage, Beverly Hills, 1980.
5. Morrison, D.F., *Multivariate Statistical Methods*, McGraw Hill, London, 1990.