Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Jastrzębia Góra, 16th–20th September 2013

APPLYING DATA MINING CLASSIFICATION TECHNIQUES TO SPEAKER IDENTIFICATION

Kinga Sałapa[1,†], Agata Trawińska[2] and Irena Roterman-Konieczna[1,*]

[1]Department of Bioinformatics and Telemedicine,
Jagiellonian University Medical College,
ul. św. Łazarza 16, 31-530 Kraków, Poland
[2]Institute of Forensic Research,
ul. Westerplatte 9, 31-033 Kraków, Poland
[†]kinga.salapa@uj.edu.pl, [2]atrawinska@ies.krakow.pl,
[*]irena.roterman-konieczna@uj.edu.pl

ABSTRACT

Voice is one of biometric measure which can characterize an individual as unique in the whole world. Unfortunately, this assumption has not been proven so far and treating speech signal as DNA or fingerprints is not relevant. The researchers from many forensic disciplines try to find the best both signal acoustics feature(-s) and model(-s) to distinguish people via their voices. The goal of this investigation is to present effectiveness of Data Mining techniques to classification task. Four algorithms were applied such as C&RT and CHAID classification trees and MLP and RBF neural networks models. The results show their high force to distinguish speaker. It is likely that their strength lies in ability to learn complex, nonlinear relations hidden in input data without any assumptions of data and model.

INTRODUCTION

Speech is a fundamental part of communication process in everyday life of millions of people all around the world [1]. Identifying familiar people by their voices seems to be natural for human being. It seems reasonable to assume that voices are unique, but this has not been scientifically proven [2, 3]. The complexity of this problem is emphasized by many researchers. Voice pattern is one of several types of biometric characteristics of an individual and biometric methods have become one of the most convincing ways to confirm the identity of the individual [4]. The interpretation of recorded speech as biometric evidence in forensic context presents particular challenges [5].

Phoneticians are able to measure features of the acoustic speech signal, but it is still not know a set of criteria by which the voices of individuals can be distinguished uniquely. Recent studies concerning formant frequency analysis indicate that this approach may provide valuable clues [6, 7]. These studies differ with respect to the statistical tools used to assess the accuracy of speaker identification. They vary from simplest statistics to multidimensional analysis such as MANOVA or, the most popular, discriminant analysis. Unfortunately, these analyses come with a number of preconditions [8]. These preconditions significantly restrict practical applicability of multidimensional analysis methods. Data Mining techniques are very attractive analyses alternatives.

Their fundamental advantage is the complete lack of restrictions regarding input data and classification models [9]. Neural networks, along with other Data Mining paradigms such as Hidden Markov Model (HMM), Support Vector Machine (SVM) or Guasian Mixture Model (GMM), are frequently applied in Automatic Speaker Verification systems [10].

The goal of present investigation is to compare effectiveness of some of Data Mining techniques such as neural network and classification trees in speaker identification. Their task is to decide, who among many candidates (speakers) said it, given a sample of speech. Hence, this is an N–class decision task, where N is the number of speakers.

MATERIALS

The study is based on recordings of ten sentences, each voiced three times by five males representing the Lesser Polish dialect, aged 21–23 (denoted: S1–S5). Subject were recorded in the sound–treated room in the lab at the Institute of Forensic Research (Cracow, Poland). Recordings were obtained in lossless WAV PCM format, with a sampling rate of 44.1 kHz and 16–bit sample resolution. This paper only presents results obtained for a subset of acoustic realizations of the a vowel, *i.e.* two repetitions of unstressed a from the following contexts: $p - a$ and $n - a - l$, described in terms of the lowest four formants (F1–F4). Formant frequencies were extracted automatically using the STx software tool published by the Austrian Academy of Sciences.

METHODS

The main part of this investigation (*i.e.* construction of classification models) was preceded by input data pre–processing, which focuses on searching for univariate (z–score) and multivariate (Mahalanobis D2 metric) outliers. A data point was considered as univariate outlier if z–score was above 2.5, and as multivariate outlier if the probability associated with its Mahalanobis' distance was 0.001 or less [9].

Four classification models were constructed for each context. The first two of them focused on an application of classification tree algorithms to detect criteria for dividing the whole datasets into five determined by speakers classes. The two types of classification trees were used, *i.e.* C&RT (*Classification and Regression Trees*) and CHAID (*Chi–squared Automatic Interaction Detector*). It can be pointed out two basic differences between these algorithms. The first one concerns on splitting criteria applied to make the best separation of each node. C&RT uses the Gini index while CHAID uses chi–squared test. Moreover, C&RT model is always binary *i.e.* each node can be split into two child nodes only, like it is shows on Fig. 1. That restriction does not concern CHAID trees [9, 11]. Moreover, v–fold cross–validation were applied (with v=15) to prevent overfitting the data and to be able to generalize the models for new items.

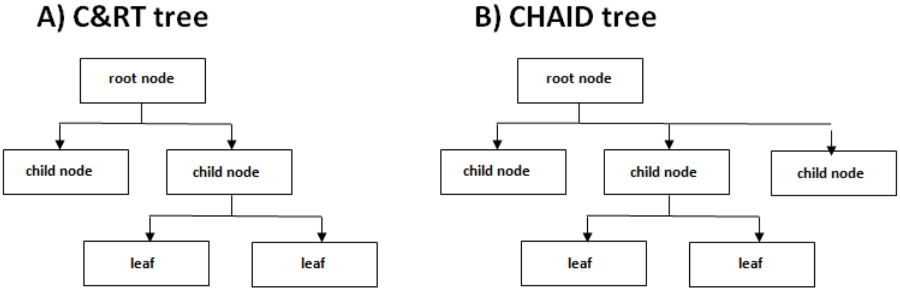

Figure 1. Topology of C&RT and CHAID classification trees.

The other two types of model were based on artificial neural network. In classification–oriented tasks the most frequently used type of neural network is the feed–forward model, which can be further divided into *Multilayer Perceptron* (MLP) and *Radial Basis Function* (RBF) networks. Both represent the supervised learning approach, in which each class is defined by the researcher [9]. Artificial neural networks mimic the operation of biological neurons in the human brain. They emulate the brain's complexity (collectivism) and its adaptation to various types of data [12]. Prior to analysis input data was randomly divided into training sample, testing sample and validating sample (using a population ratio of 0.70:0.15:0.15) in order to avoid excessive adaptation of the model to empirical data. The role of the training set is to adjust input weights; the test set enables on–the–fly monitoring of the training process, while the validation set can be used to assess the final outcome of training. The search for an optimal neural network model was based on an automatic network designer. For each type of network 500 classification models were constructed and from this group one model was ultimately selected, based on accuracy and consistency of results obtained for each dataset. The models contained a single hidden layer with not more than 50 neurons (Fig. 2).

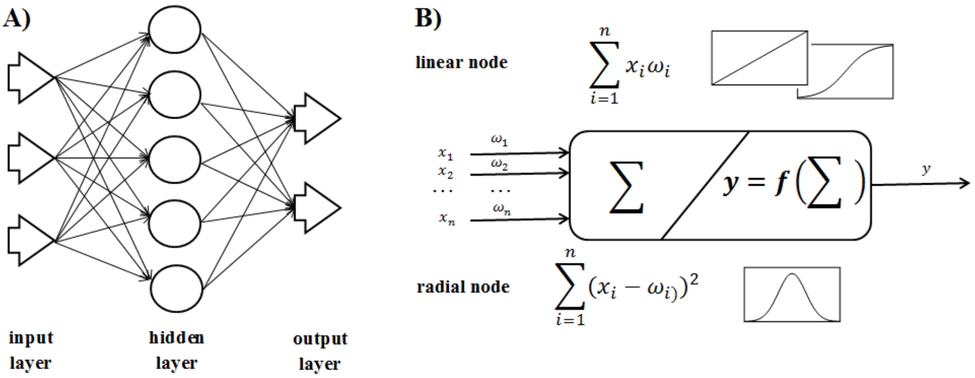

Figure 2. Architecture of the neural network and a single node.

Final assessment of model quality was based on correct classifications obtained using the validation set in case of neural network models and on results of v–fold cross–validation in case of classification trees. The percentage of correct classifications (for a given context) was compared using chi–square test. When the number of theoretical samples was limited Yates's correction was carried out. Furthermore, in case of statistically significant results were obtained comparing more than two proportion algorithm was applied [13].

Results were deemed statistically significant when the calculated p–value did not exceed the statistical significance threshold ($\alpha = 0.05$). All computations were carried out using the *STATISTICA Data Miner* software (StatSoft Inc., Tulsa, OK, USA). Additionally, the study involved a set of macros written in *STATISTICA Visual Basic*.

RESULTS

This investigation shows high efficiency of Data Mining techniques as a speaker identification tools. Percentages of correct identifications of speakers whose speech was subjected to analysis considerably exceed random classification results in all models. Moreover, models based on $p - a$ context were perfectly accurate in six cases including three MLP neural and two RBF neural network models, and one C&RT classification tree (Fig. 3). All models consisted of at least three formant frequencies.

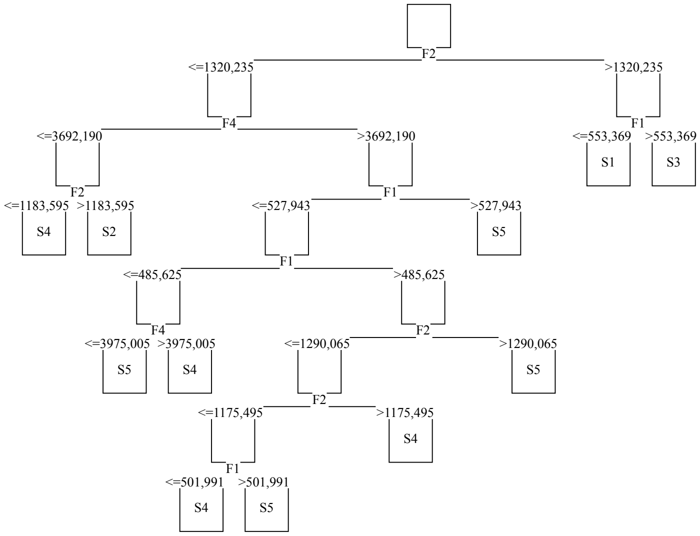

Figure 3. C&RT trees for $p - a$ context concerning F1, F2 and F4.

The task of classification trees is to find the path of dividing data sets into several groups. According to Fig. 3, presenting the best C&RT model for $p - a$ context and based on tree formants such as F1, F2 and F4, database is firstly divided by the second formant. The threshold value is equal to 1320,235 Hz. If frequency of this formant for an item is above it the item can be the first (S1) or the third speaker (S3), depending on its first formant frequency. If its value is above 553,369 Hz an item will be classified as S3, otherwise as S1. On the other hand, please notice that there is several different paths to classify an item as S4 as well as S5. For example, the latter speaker can be recognized if its second formant is lower than 1320,235 Hz, its fourth formant is above 3692,190 Hz and its first formant exceeds 527,943 Hz. But it also can be recognized as S5 if: F2 <1320,235 Hz, 3975,005>F4>3692,190 Hz and F1<485,625 Hz.

Fig. 4 presents the rates of positive classifications (in percentages) according to applied all possible combinations of predictors (*i.e.* formant frequencies) and for all applied techniques. Three conclusions deserve special attention. Firstly, C&RT classification trees gave the best results in almost all models specially when more than one formant contour were used as predictors. Secondly, CHAID classification trees had the worst results in almost all models. And finally, it should be pointed out that results obtained from both neural networks models (MLP and RBF) are nearly identical.

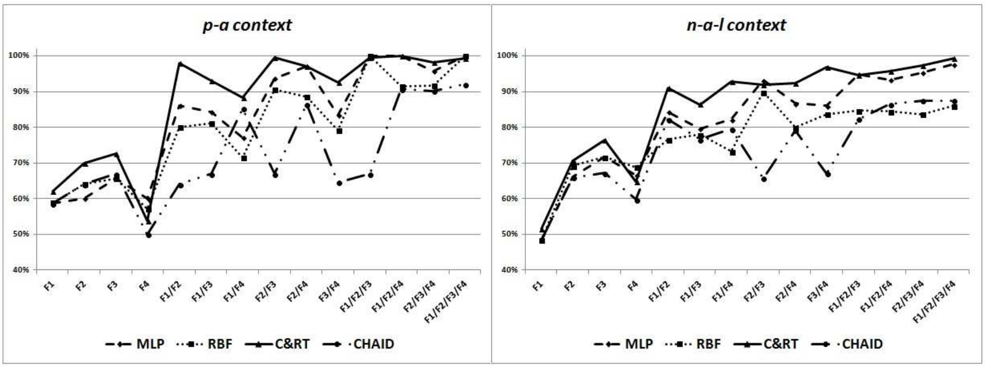

Figure 4. Percentage of positive speaker identification.

Table 1 shows the exact p–value of chi–square test or Yates correction as well as results of comparing each pair of rates. The latter confirm statistically observations presented in the previous paragraph.

Table 1. Comparison of percentage of correct speaker classification

	p–a context		*n–a–l context*	
	p-value	multiple comparison	p-value	multiple comparison
F1	0.8074	—	0.8031	—
F2	0.2834	—	0.5168	—
F3	0.5651	—	0.0384	C&RT vs. CHAID
F4	0.6078	—	0.4436	—
F1/F2	0.0000	MLP vs. C&RT MLP vs. CHAID RBF vs. C&RT RBF vs. CHAID C&RT vs. CHAID	0.0003	RBF vs. C&RT C&RT vs. CHAID
F1/F3	0.0000	C&RT vs. CHAID	0.0045	C&RT vs. CHAID
F1/F4	0.0322	RBF vs. C&RT	0.0000	RBF vs. C&RT C&RT vs. CHAID
F2/F3	0.0000*	RBF vs. C&RT C&RT vs. CHAID	0.0000	MLP vs. CHAID RBF vs. CHAID C&RT vs. CHAID
F2/F4	0,0006*	C&RT vs. CHAID	0.0001	RBF vs. C&RT C&RT vs. CHAID
F3/F4	0.0000	C&RT vs. CHAID	0.0000	MLP vs. C&RT MLP vs. CHAID RBF vs. C&RT RBF vs. CHAID C&RT vs. CHAID
F1/F2/F3	0.0000*	MLP vs. CHAID RBF vs. CHAID C&RT vs. CHAID	0.0000	MLP vs. CHAID RBF vs. C&RT C&RT vs. CHAID
F1/F2/F4	0.0001*	MLP vs. C&RT RBF vs. C&RT &RT vs. CHAID	0.0011*	RBF vs. C&RT &RT vs. CHAID
F2/F3/F4	0.0463*	C&RT vs. CHAID	0.0001*	RBF vs. C&RT C&RT vs. CHAID
F1/F2/F3/F4	0.0122*	C&RT vs. CHAID	0.0000*	RBF vs. C&RT C&RT vs. CHAID

* Yates correction, otherwise chi–square test

ACKNOWLEDGEMENTS

I would like to thank colleagues from the Institute of Forensic Research for speech recordings' sharing.

REFERENCES

[1] M. Tiwari and M. Tiwari: *Voice—How humans communicate?*, J. Nat. Sc. Biol. Med. **3** (2011), 3–11.

[2] A. Drygajło: *Statistical evaluation of biometric evidence in forensic automatic speaker recognition*, Proceedings of Computational Forensics, Berlin Heidelberg, 2009, pp. 1–12.

[3] A. Trawińska and A. Klus: *Forensic speaker identification by the linguistic-acoustic method in KEU and IES*, Problems of Forensic Sciences **LXXVIII** (2009), 160–174.

[4] A. Jain: *Handbook of Biometrics*, Springer, New York, 2008.

[5] C. Champod and D. Meuwly: *The inference of identiry in forensic speaker identification*, Speech Commun. **31** (2000), 193–203.

[6] F. Nolan: *Speaker identification evidence: its forms, limitations and roles*, Proceeding of the conference Law and Language: Prospect and Retrospect, Houston, 2001, pp. 1–19.

[7] K. McDougall: *Dynamic features of speech and the charakterization of speakers: towards a new approach using formant frequencies*, Int. J. Speech Lang. La. **13** (2006), 89–126.

[8] B.G. Tabachnikand and L.S. Fidell: *Using Multivariate Statistics*, Pearson, New York, 2012.

[9] S. Tufféry: *Data Mining and Statistics for Decision Making*, Wiley, Hoboken, 2011.

[10] T. Kinnunen and H. Li: *An overview of text–independent speaker recognition: From features to supervectors*, Speech Commun. **52** (2010), 12–40.

[11] K. Sałapa, B. Kalinowska, T. Jadczyk, and I. Roterman: *Measurement of Hydrophobicity Distribution in Proteins—Complete Redundant Protein Data Bank*, Bio-Algorithms and Med-System **8** (2012), 195–206.

[12] W. Jassem and W. Grygiel: *Off–line classification of Polish vowel spectra using artificial neural networks*, J. Int. Phon. Assoc. **34** (2004), 37–52.

[13] J.H. Zar: *Biostatistical Analysis*, Pearson, New York, 2010.