DOI: 10.21008/j.0860-6897.2022.2.03
Vibrations in Physical Systems, 2022, 33(1), 2022203

Recognition of Speaker's Age Group and Gender for a Large Database of Telephone-Recorded Voices

Piotr STARONIEWICZ

Wrocław University of Science and Technology, Faculty of Electronics, Photonics and Microsystems, Department of Acoustics, Multimedia and Signal Processing, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

Corresponding author: Piotr STARONIEWICZ, email: piotr.staroniewicz@pwr.edu.pl

Abstract The paper presents the results of the automatic recognition of age group and gender of speakers performed for the large SpeechDAT(E) acoustic database for the Polish language, containing recordings of 1000 speakers (486 males/514 females) aged 12 to 73, recorded in telephone conditions. Three age groups were recognised for each gender. Mel Frequency Cepstral Coefficients (MFCC) were used to describe the recognized signals parametrically. Among the classification methods tested in this study, the best results were obtained for the SVM (Support Vector Machines) method.

Keywords: speech processing, automatic age recognition.

1. Introduction

The human voice provides information not only about the content of the utterance (semantics), but also a whole range of non-verbal information, such as personal characteristics, health status, fatigue, emotional state, gender or age of the speaker. A still under-researched and under-appreciated aspect of voice recognition techniques is the determination of the speaker's age. One reason for this is the relatively small number of acoustic databases containing accurate information about the age of speakers. Information about the speaker's age can be particularly important during telephone conversations (when we cannot judge age from appearance). Today's advertising offers and services are tailored to the age of the customer and such information may be relevant to, for example, tele-interviewers (who now often use software that recognises the emotional state of the caller). Another important application is in forensics, where, at the investigation stage, it is important to determine the age range when creating a criminal profile from audio recordings. The results of automatic speaker age recognition obtained in studies to date show very large differences. For the most part, they have been carried out on publicly available, free corpora of recordings such as TIMIT or Common Voice for English, which is a great advantage as it allows the search for the most effective algorithms and their comparison between different research centres [1]. At the same time, tests conducted on other corpora or for other languages have shown that the results obtained can vary significantly depending on the acoustic database used (e.g. significant differences obtained for the two telephone databases compared for German [2]). In addition to the quality of the signal (determined by the coding methods used in telephony and the transmission conditions), such recordings have their own characteristics, especially when it is a conversation with a tele-interviewer or a stranger, and, as mentioned earlier, this is one of the main potential applications of automatic age recognition systems. During such a telephone call statements are generally laconic (greetings, introduction) and interviewees often use short answers ("yes"/"no"). The shorter-than-usual duration of utterances and the limited semantic content of the test material may affect the performance of text-independent systems, such as voice verification systems or, finally, speaker age recognition. In this study, tests were carried out for a relatively large telephone database for the Polish language (described further in the paper) and for typical, not very extensive parameterisation and classification methods used in voice recognition in order to be able to compare in the future with results obtained for other databases.

2. Polish SpeechDat(E) Database

The SpeechDat(E) database (Eastern European Speech Databases for Creation of Voice Driven Teleservices) for Polish was recorded as part of the INCO-Copernicus project No. 977017 at the Wrocław University of Science and Technology, and is distributed by ELRA (European Language Recources Association) [3]. The Polish SpeechDat(E) database has been assigned the ISLRN: 748-036-853-302-2 (International

Standard Language Recourse Number). All speech databases produced as part of the SpeechDat(E) project were validated by SPEX, the Netherlands, which assessed their compliance with the SpeechDat(E) format and content specifications. In terms of the number of voices, this database is still one of the most extensive acoustic databases available for the Polish language. The database contains recordings of 1,000 people recorded over a fixed telephone network using an ISDN interface and is provided on 5 CD-ROMs recorded in ISO9660 standard with 200 speakers on each disc. The audio files were recorded according to the ITU-T recommendation for G.711 PCM coding (Pulse Code Modulation) as 8bit, 8kHz sampling with A-law coding.

Table 1. Items recorded by each speaker in the Polish SpeechDat(E) database.

Recorded items		Number of items
Application words		6
Sequence of 10 isolated digits		1
Connected digits	sheet number	1
	telephone number	1
	credit card number	1
	PIN code	1
Dates	spontaneous date	1
	prompted date	1
	relative and general date expression	1
Word spotting phrases		1
Isolated digits		1
Spelled words/phrases	spontaneous (forename)	1
	directory assistance name	1
	artificial for coverage	1
Currency amounts	amount in Polish zloty (PLN)	1
	amount in US dollars or euros (USD or EUR)	1
Natural numbers		1
Directory assistance names	spontaneous (forename)	1
	spontaneous (city of growing up)	1
	city name (set of 500 cities)	1
	company/institution name (set of 500)	1
	forename and surname (set of 150)	1
	forename (set of 150)	1
Questions	predominantly ”yes” question	1
	predominantly ”no” question	1
Phonetically rich sentence	(set of 1,536)	12
Time phrases	spontaneous time of day	1
	time phrase (word style)	1
Phonetically rich word	(set of 1,320)	4
	Total	48

Most items in the database (see Tab. 1) contain acoustic material for teleservice applications [4]. The exception is a collection of 12 phonetically rich sentences spoken by each speaker. The sentences were up to 9 words long and came from a variety of texts, such as books and newspapers. The Polish corpus of 1,536 sentences was collected with special care for a good coverage of rare phonemes. The entire corpus was divided into twelve collections, each containing a specific group of the rarest phonemes. Finally, the division into sets of 12 phonetically rich sentences was done by a computer program (a different set for each speaker in the database), in such a way that each set contained at least 2 examples of each Polish phoneme. This resulted in 1280 sets of sentences.

According to the assumptions made for the database, the recordings were made all over the country via telephone handsets of various types (almost 10% of the calls were recorded using outdated telephone handsets equipped with low-quality carbon microphones). Also, almost 30% of the total number of recordings were calls made from public places with high noise levels. Despite such difficult conditions, the majority of the recordings are of good or very good quality sufficient for use in automatic speech recognition systems. This was confirmed by both objective measurements and subjective evaluation, which was carried out for the whole of the recordings by the performers of the database and for a part of it by SPEX. In order to determine the quality of the signal in the base, the distributions of the values of three parameters were determined: clipping ratio, signal-to-noise ratio and mean value of the samples (see Eqs. (1)-(3) and Figs. 1-3).

The clipping ratio (C_R) is defined as the ratio of the number of samples in the file that have a maximum or minimum value (L_m) to the total number of samples (L_p):

$$C_R = \frac{L_m}{L_p} \tag{1}$$

The mean value of the samples in the signal can be used to detect recordings with a large DC-offset and was defined as follows (where: V_i is the value of the ith sample of the signal and L_p is the total number of samples):

$$M_V = \frac{\sum_{i=1}^{L_p} V_i}{L_p} \tag{2}$$

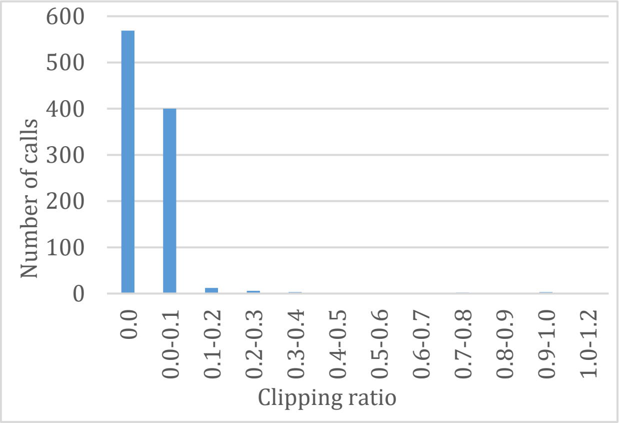

Figure 1. Histogram of the number of calls in function of the clipping ratio value.

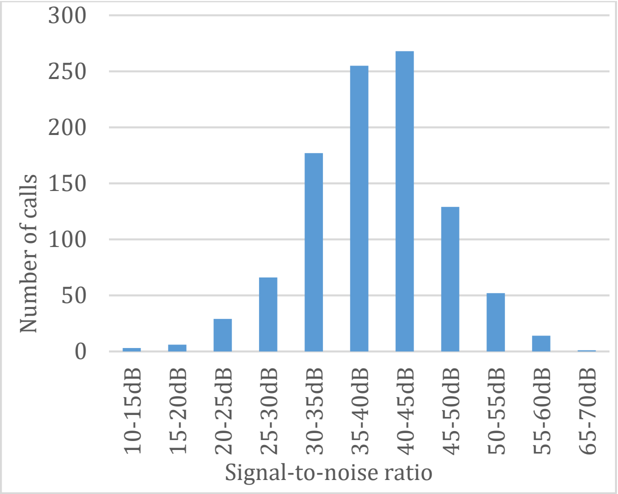

Figure 2. Histogram of the number of calls in function of the signal-to-noise ratio value.

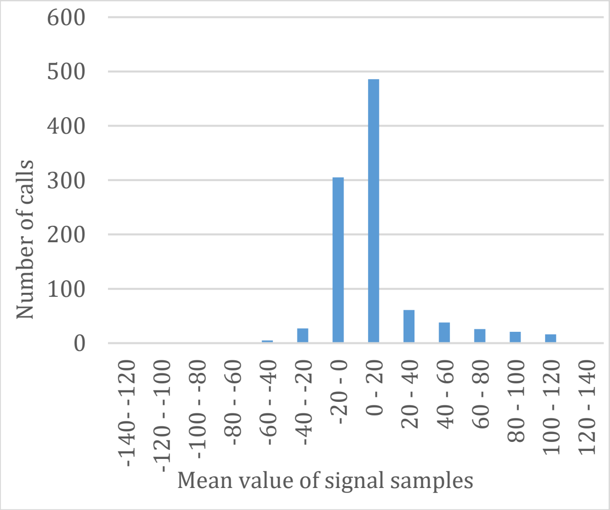

Figure 3. Histogram of the number of calls in function of the mean value of the signal samples.

Due to the high quality and versatility of the recorded utterances, the database can be used in speech and speaker recognition research, with a particular focus on the recently increasingly popular teleservice applications (e.g. telephone booking, tourist information, telephone banking orders, etc.). An additional advantage is that analogous material recorded according to the same technical requirements has already been recorded for most European languages. This simplifies the process of developing and implementing already existing systems for other languages.

3. Experiments and discussion

The database consisted of recordings of speakers between the ages of twelve and seventy-three. During the recording, each speaker provided a scroll of their age. Fig. 4 shows the number of speakers appearing in the database for a particular age. As can be seen in Fig. 4, the database is quite well-balanced in terms of the speakers' age and has a good representation of voices from all age groups. The requirements of the SpeechDat(E) project called for a mandatory representation of at least 20% of the recordings in the three age groups: 16-30 years, 31-45 years and 46-60 years, which was strictly adhered to.

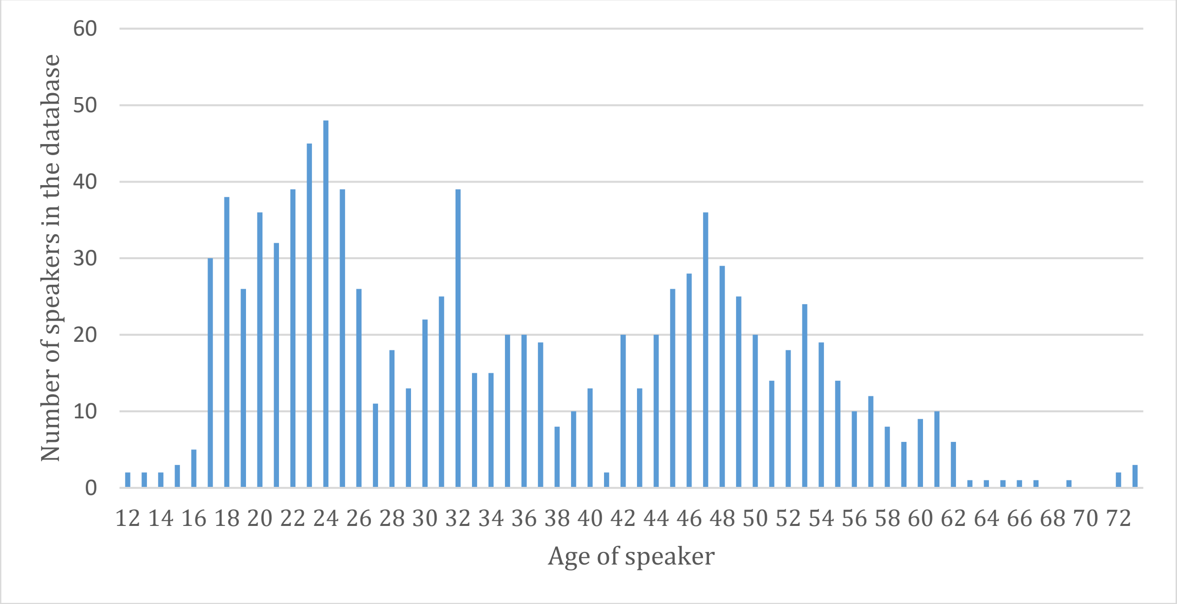

Figure 4. Distribution of the number of speakers in the database according to their age.

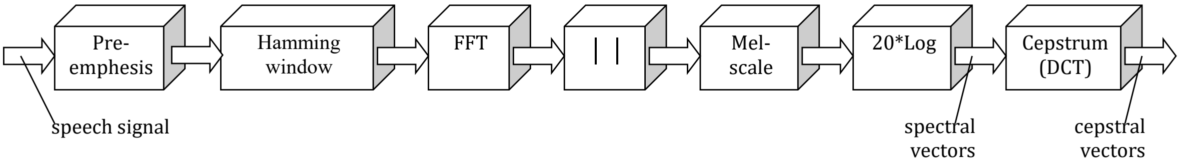

Figure 5. Procedure of MFCC.

It was decided to divide the speakers into three age groups with similar numbers. Those aged between twelve and twenty-five (347 speakers) were allocated to the first age group, the second group was between twenty-six and forty-three (327 speakers) and the third group between forty-four and seventy-three (326 speakers). An additional criterion for division was the gender of the speakers. Based on this characteristic, a division was made into two groups, where the first group was male (486 speakers) and the second group was female (514 speakers).

The MFCC parameters (Mel-Frequency Cepstral Coefficients) are among the most commonly used parameters in speech and voice recognition. They are determined according to the scheme shown in Fig. 5.

In the process of obtaining the MFCC parameters, the first step is to divide the signal into frames of a specific length and offsets between adjacent windows. The fast Fourier transform (FFT) and modulus are then determined. The next step is the filtering process, for which band-pass filters are used, after which the logarithm of the results obtained is determined, which is then multiplied by 20 to obtain the spectral vectors on a decibel scale. In the final step, a cosine transformation is performed [5].

For the parameter vectors thus obtained, a learning and classification process was carried out using three methods: Random Forest (RF), Support Vector Machines (SVM) and Multilayer Perceptron (MLP). The Random Forest algorithm is a machine learning algorithm based on the concept of combining decision trees into one large set, or forest. In a random forest classification, each tree decides whether a sample belongs to one of the classes, and then, as a result of a majority vote, this sample is classified [6]. The idea behind SVM classification is to map the data in space and then determine a plane separating the vectors belonging to each class [7]. For a linear SVM classifier, a hyperplane has to be determined that separates the classes of the training set by as large a margin as possible. The samples of the test set are classified based on their location with respect to the determined hyperplane. SVMs in classifying non-linearly separable data often use sets of mathematical functions, which are called kernels. Their use is to transform samples of a training set into a desired form, so that classification can be performed. One example of such a kernel is, for example, the polynomial kernel and also the Gaussian RBF (Radial Base Function) kernel used in this paper. Neural networks are powerful machine learning tools that have high accuracy and the ability to accommodate multiple data. Among the most commonly used networks is the multilayer network, which is built with an input layer, at least one hidden layer and an output layer. In the input layer, data is loaded, which then goes to the first hidden layer, where the learning process takes place. The task of the output layer is to compute the entire network and then return the output data. Different types of neural networks are distinguished by their architecture and include recurrent networks, convolutional networks or MLP (Multilayer Perceptron) networks. Table 2 shows the recognition results for the three classification methods tested. For reference, the values in addition to those obtained by the automatic methods are also given in the table for auditory recognition. Forty-five listeners took part in subjective listening tests. For listening and classification into three age groups, samples were selected from the same SpeechDat(E) database for Polish: 15 male voice samples and 15 female voice samples for each of the three age groups, with an even age distribution of speakers.

Table 2. Recognition results for the three classification methods tested and humans.

	Males	Females	Jointly	Gender
RF	62%	43%	54%	77%
SVM	83%	80%	75%	90%
MLP	68%	70%	68%	86%
Humans	56%	67%	61%	-

4. Conclusions

This paper presents the results of automatic recognition of the speakers' age group and gender for the Polish language database recorded under conditions of telephone transmission, i.e. the most likely practical application for this type of task. Among the classification methods tested in this study, the best results were obtained for the SVM method which is supported by the previous work. Interestingly, the listening subjective results turned out to be much weaker and with a much greater disparity in age recognition between men and women than is the case for the automatic methods, which may be an interesting research aspect for further studies.

Despite the text-independent systems used, the preliminary results can confirm the strong influence of the type of acoustic material on the effectiveness of such recognition (which is also confirmed by some previous papers). Future work will focus on comparative studies of the results obtained for other available acoustic databases containing other speaking styles. It is also planned to relate the results to subjective auditory tests by a group of listeners.

Additional information

The author declares: no competing financial interests and that all material taken from other sources (including their own published works) is clearly cited and that appropriate permits are obtained.

References

1. D. Kwasny, D. Hammerling; Gender and Age Estimation Metods Based on Speech Using Deep Neural Networks; Sensors 2021, 21, 4785. DOI:10.3390/s21144785
2. T. Bocklet, A. Maier, J.G. Bauer, F. Burkhardt, E. Noth; Age and gender recognition for telephone applications based on GMM supervectors and support vector machines; Proceedings of 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, USA, March 31 – April 4, 2008; IEEE: Piscataway, USA, 2008. DOI: 10.1109/ICASSP.2008.4517932
3. P. Pollak, J. Cernocky, J. Boudy, K. Choukri, H. Heuvel, K. Vicsi, A. Virag, R. Siemund, W. Majewski, J. Sadowski, P. Staroniewicz, H. Tropf, J. Kochanina, A. Ostrukhov, M. Rusko, M. Trnka; SpeechDat(E) – eastern European telephone speech databases; In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00); Athens, Greece, May 31 – June 2, 2000; European Language Resources Association: Athens, Greece, 2000.
4. P. Staroniewicz, J. Sadowski; SpeechDat Polish Database for the Fixed Telephone Network, (Polish Database documentation file), 2000 (http://www.elra.info).
5. F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delcretaz, D.A. Reynolds; A Tutorial on Text-Independent Speaker Verification; EURASIP J. Adv. Signal Process. 2004, 101962. DOI:10.1155/S1110865704310024
6. T.K. Ho; Random decision forests; Proceeding of 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, August 14-16, 1995; IEEE: Piscataway, USA, 1995. DOI:10.1109/ICDAR.1995.598994
7. Y.W. Chang, C.J. Hsieh, K.W. Chang, M. Riggaard, C.J. Lin; Training and Testin Low-degree Polynomial Data Mappings via Linear SVM; Journal of Machine Learning Research 2010, 11(48), 1471–1490.