Speech Synthesis for Gender Classification

https://doi.org/10.3991/ijes.v5i1.6690

Kawther A. Al-Dhlan
University of Hail, Hail, Kingdom of Saudi Arabia
kawthar.al.dahlan@gmail.com

Abstract—This paper presents a gender identification system to be used for call forwarding in health related communications. The system listens to the caller then using speech synthesis, image processing, and linear support vector machine SVM identifies either he or she is a male or a female. This solution is imperative in a conservative country such as the Kingdom of Saudi Arabia in order to forward the call to a male or female practitioner. The originality of the approach is that no transcription is used to learn SVM models. To identify the gender of the caller, the trained SVM model of the reference pieces are compared to transcripts of the audio frequency record and are using the Levenshtein distance. For the identification of gender, we obtain an accuracy rate of 94% on a test flow containing 449 pieces of speech clips.

Keywords—Linear SVM; Machine Learning ;spectrogram

1 Introduction

Audio content identification consists of retrieving metadata (artist, album name, song name, advertising name, etc.) from an unknown audio clip. There are many potential applications for audio identification, the most popular being automatic radio stream monitoring and the identification of an unknown audio clip captured by a mobile device. Manually performing the task of audio identification is rather tedious and slow. To address this problem, there are two main approaches: audio tattooing and fingerprint extraction. The audio tattoo consists of hiding the information to be identified (artist name, album ...) in the audio document.

The aim of this approach is to inject the desired information without altering the audio quality of the document. In the next step, a signature is extracted from the unknown content and compared to the references' fingerprints stored in a database. An acoustic fingerprint is a compact presentation of the audio content. We are interested in methods based on the extraction of audio fingerprints, which are more suitable for the automatic monitoring of radio broadcast. The audio identification by fingerprint extraction consists of two modules: a fingerprint extraction module and a comparison module. The first step in an audio identification system based on fingerprint extraction is the creation of a fingerprint base from a reference database. The reference database contains the audio documents (music, advertisements, jingles) that the sys-

tem could identify. In the second step, an unknown audio clip is identified by comparing its fingerprint with those of the reference database.

A soundtrack is a compact presentation of the audio content. We are interested in methods based on the extraction of acoustic fingerprints, which are more suitable for the automatic monitoring of radio broadcast and automatic call forwarding. The automated acoustic identification by fingerprint extraction consists of two modules: a fingerprint extraction module and a comparison module. The first step in an acoustic identification system based on fingerprint extraction is the creation of a fingerprint base from a reference database. The reference database contains the sound files (speeches, music, advertisements, and jingles) that the system could identify. In the second step, an unknown sound clip is identified by comparing its fingerprint with those of the reference database. Acoustic identification by fingerprint extraction has been extensively studied over the past several years. Thus, the state of art is relatively advanced, with proposals for very diverse approaches to address the problem [1]. The main thrust of these systems is to compute a robust acoustic fingerprint against different types of distortions and to propose a rapid method of comparison which can satisfy the real-time constraints. In this paper, we present a system that is able to identify the gender of a caller in an initiated phone call. Based on the extraction of the local frequency peaks of the acoustic signal spectrogram, our technique makes it possible to achieve greater robustness to the distortions present in a phone call.

This paper is structured as follows: Section 2 presents a state of the art of mainstream audio extractions by fingerprint extraction. Section 3 presents our audio identification system based on the ALISP tools. Section 4 describes the adopted experimental protocols and sets out the results obtained for the identification of a caller's gender in a phone call.

2 Acoustic Identification Systems Based on Fingerprint Extraction

As shown in Fig. 1, the general architecture of a fingerprint identification system is based on a learning module which registers fingerprint references into a database and then an identification module compares an unknown sound file fingerprint against the registered references fingerprints. Several methods of audio identification by fingerprint extraction have been proposed [1]. We have chosen to present these systems according to the approach used for fingerprint extraction. Through the papers published on the subject, three major families emerge with regard to the technique of fingerprint extraction. The first family operates directly on the spectral representation of the signal to extract the fingerprints. This type of fingerprint is generally easy to extract and does not require significant computing resources. The second family uses the techniques used in the field of computer vision; the main idea is to treat the spectrogram of each sound file as a 2-D image and to transform the acoustic identification into an image identification problem, which includes approaches based on vector quantization and machine learning, these systems offer an impression model that imitates techniques used in speech processing.

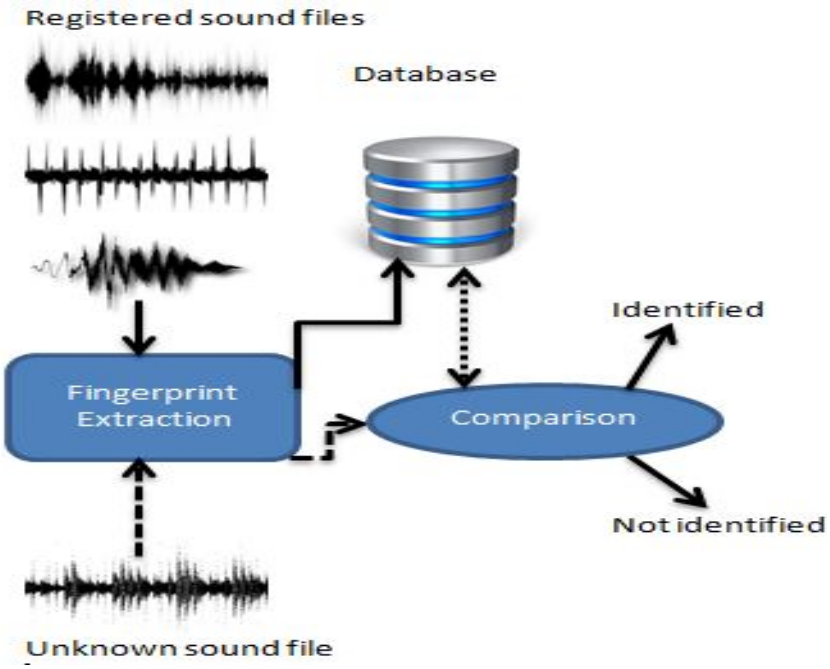

Fig. 1. General architecture of an acoustic identification system based on fingerprinting

2.1 Spectral Representation Based Techniques

These techniques are most commonly used for the simplicity of fingerprint extraction. Several systems use the spectral representation of the signal directly to construct the fingerprint Fig. 2 shows the logarithmic plot of a sound file spectrogram's magnitude. Haitsmaetal [2] have developed an audio identification system for the recognition of music tracks (method proposed by Philips). They used a Bark scale to reduce the number of frequency bands through 33 logarithmic bands covering the 300Hz to 2kHz range. The sign of the energy difference of the adjacent bands is calculated and stored in binary form. The result of this quantization process is a 32-bit fingerprint per frame. The search method adopted by Philips consists in indexing each reference frame in a lookup table. If the number of sub bands used is N_b, then each frame will be represented by a vector of (N_{b-1}) bits and the lookup table will then have entries. The binary bitmap of the fingerprint serves as the key in the lookup table, with all the reference fingerprints having the same binary pattern of a fingerprint to be identified running as candidates for identification. Haitsma et al. assumed that there exists at least one binary frame of the imprint to be identified which is not distorted with respect to the corresponding reference. This technique has given rise to various studies. Y. Liu [3] modified the algorithm to bypass the hypothesis of having an undistorted binary frame while the authors themselves tried to improve the extraction path method in order to make the system more robust to distortions such as the temporal stretching (pitching) [4].

Another commercial system (Shazam) based on the spectral representation of the system was proposed by Wang [5] for the identification of an unknown audio extract captured by a mobile phone. This technique transforms the spectrogram into a binary image by keeping only local maxima Fig. 2. It is then necessary to extract peaks from this spectrogram, taking care to choose points of maximum energy locally and ensuring a homogeneous peak density within the spectrogram. The author proposes to index the characteristics of the references using the localization of the peaks as index. However, an index based on the location of each point in isolation is not very selective. Therefore, Wang proposes to use pairs of peaks, where each peak being combined with its nearest neighbors. This technique is used to identify a piece of music in a noisy environment.

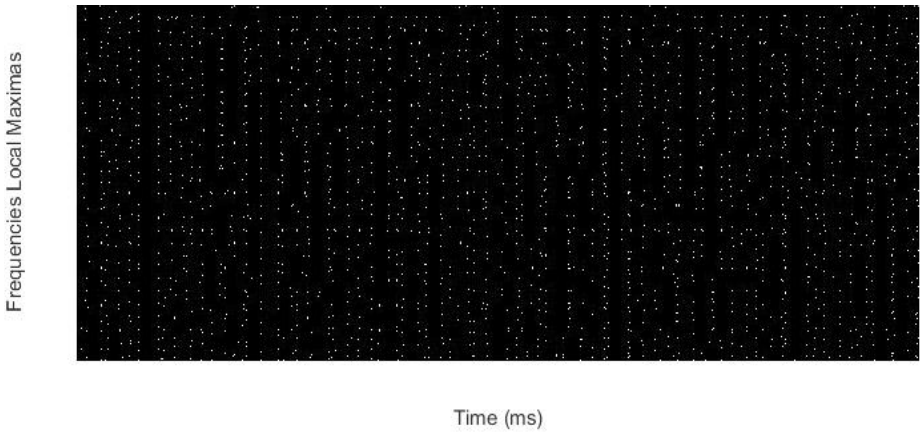

Fig. 2. Frequency's local maxima of a spectrogram

2.2 Image Processing Based Techniques

There have been several experiments in the use of image processing techniques for audio identification by fingerprint extraction. The main idea is to treat the spectrogram of each audio document as a 2-D image. Baluja et al. [6] have exploited the applicability of wavelets in the search of images in large databases to develop an audio identification system using fingerprint extraction. This technique consists in generating a spectrogram from an audio signal using the same procedures as [2], resulting in 32 bands of energetic logarithm between 318Hz and 2kHz for each frame. Then, a spectral image is extracted from the combination of the energy bands on a certain number of frames and the wavelet decomposition using the Haar wavelets is applied to the obtained images. The sign of the first 200 wavelets' amplitudes are held in the fingerprint. Finally, a hacking set is used to find the best fingerprints and the Hamming distance is calculated between the candidate fingerprints of the soundtrack and the fingerprints of the query.

Ke et al. [7] proposed a music chip identification system based on the Viola-Jones algorithm [8]. A boosting algorithm is used on a set of Viola-Jones descriptors to

learn local and discriminant descriptors. During the learning phase, a list of candidates is determined from the descriptors previously learned. For each candidate, the RANSAC algorithm [9] is applied to align the candidate with the query and a likelihood measure is calculated between the two pieces.

2.3 Statistical Modeling Techniques

This latter family includes techniques commonly used for speech processing, such as vector quantization or hidden Markov models. Allamanche [10] proposes an approach based essentially on vector quantization. The creation of the fingerprint is done using the descriptors used in the MPEG-7 standard. The descriptors used are the intensity, the measure of spectral flatness and the spectral peak factor. The identification methodology consists in extracting these descriptors from the references, a vector quantization algorithm then produces a set of centroids (called coding vectors) approximating the vectors of the descriptors of the reference. When the system identifies an unknown extract, it extracts the signal's descriptor vectors, and then, for each reference, it projects these vectors on the coding vectors of the reference. The reference having the coding vectors which produce the minimum projection error is considered to be the reference to be identified. Cano et al. [11] proposed a system based on hidden Markov modeling. 32 HMM models called audio genes are used to segment the audio signal using the Viterbi algorithm. The audio footprint consists of sequences of labels (genes) and temporal information (time of beginning and end of each gene). During the matching process, gene sequences are extracted from a continuous radio stream and compared with the fingerprints references. In order to reduce the duration of the treatment, the DNA search algorithm called FASTA [12] was used. This system was evaluated on the task of identifying the pieces of music in a radio stream.

3 Gender Classifier Implementation

Our gender classification system is based on using the bag of visual words computer vision technique to extract the features of a male and female voice from their spectrogram images. We had two sets of voice images 51 images in the female set (set 1) and 80 images in the male set (set 2). We extracted 811520 features from set 1 and 1376896 features from set 2. Then we kept 80 percent of the strongest features from each image set, we found out that image set 1 has the least number of strongest features which was 649216. Using the strongest 649216 features from each image sets we collapsed it into 5000 words of visual features using the using k means algorithm. After identifying the feature we encoded by putting it in a specific format (histogram of features) to use it in our machine learning algorithm Fig. 3.

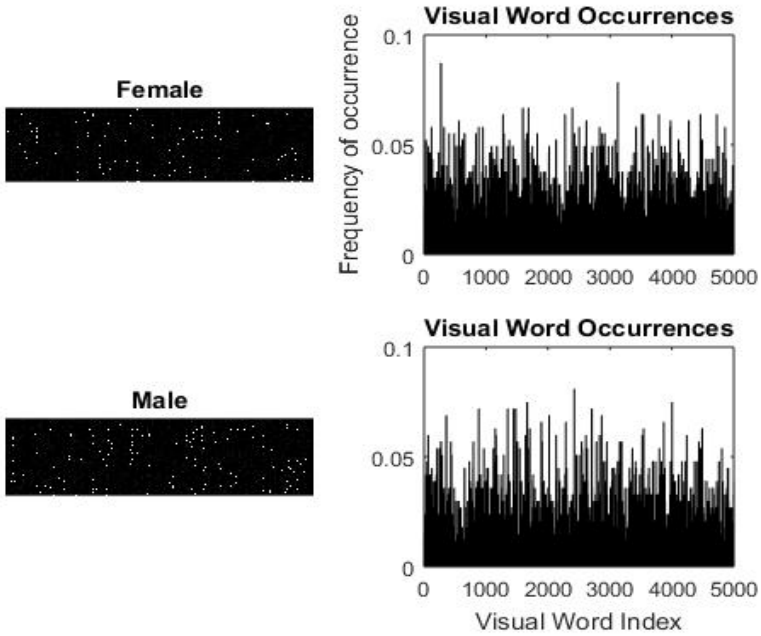

Fig. 3. Histogram of features

3.1 Learning

Machine learning uses data and produces a program to perform a task. The machine learning workflow constitutes a training phase where we preprocessed the data then we passed it into the learning phase which in our case is a classification problem and then we identified a model that worked well with our problem Fig. 4.

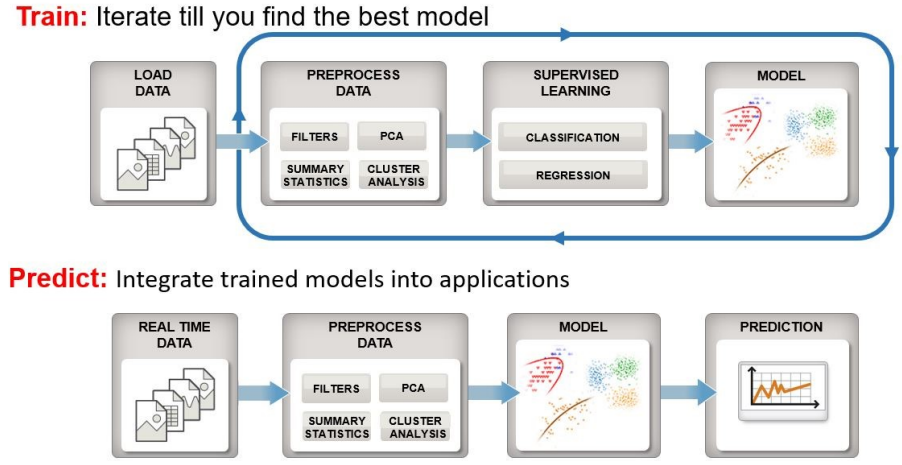

Fig. 4. Machine learning workflow

We casted the histograms of features into a table that we used in the linear support vector machine classifier which learns what each feature means. We imported the male and female speech data features with its encapsulated 5000 features and their response data column which is male of female into the classification learner. We considered different validation approaches, when we trained the data we had to set aside some data points for validation. We found out that the Linear Support Vector Machine had the best results Fig. 5 shows the confusion matrix for the linear SVM which depicts the true class and the predicted class, for the female set 46 out of 51 were correctly classified and for the male set 74 out of 80 were correctly classified.

Fig. 5. Confusion matrix

Once we have done that we went to the prediction phase where we had some new data coming (449 audio clips) in and we applied the same type of preprocessing or features extraction and then we applied the model that worked best with our data (linear SVM) which predicted based on how we trained it 94% accuracy rate.

3.2 The Database

In this study, we focused on distinguishing between the male and female voice, the 580 audio records were provided from the college of computer science and engineering at the University of Hail. We considered the demographic and linguistic backgrounds of the volunteers in order to determine which variables are key predictors of a male and female voice. The volunteers were asked to read a randomly selected phrase from a randomly selected Arabic website.

4 Results And Discussion

In this section, we present the results of the identification of female voice from male ones with our system. In order to evaluate the performance of our system, the precision (P%) measurements was used in Table 1: Accuracy: The number of female and male voice correctly detected / total number of female and male voices.

Table 1. Precision (P%) Precision: The number of female and male voice correctly detected / total number of female and male voices.

Total # of voice clips	Total # of misclassified voices	P%
449	29	94

Table 1 shows that the system was misclassified 29 voices. By analyzing these errors, and as shown in fig. 6 which depicts the histogram of features of misclassified voices "female voices classified as male voices and the other way around" we found out that the frequency of visual words occurrences for the female voices classified as male voices is almost identical to the histogram of features of a real male voice and the same goes for the a male voice classified as female voice.

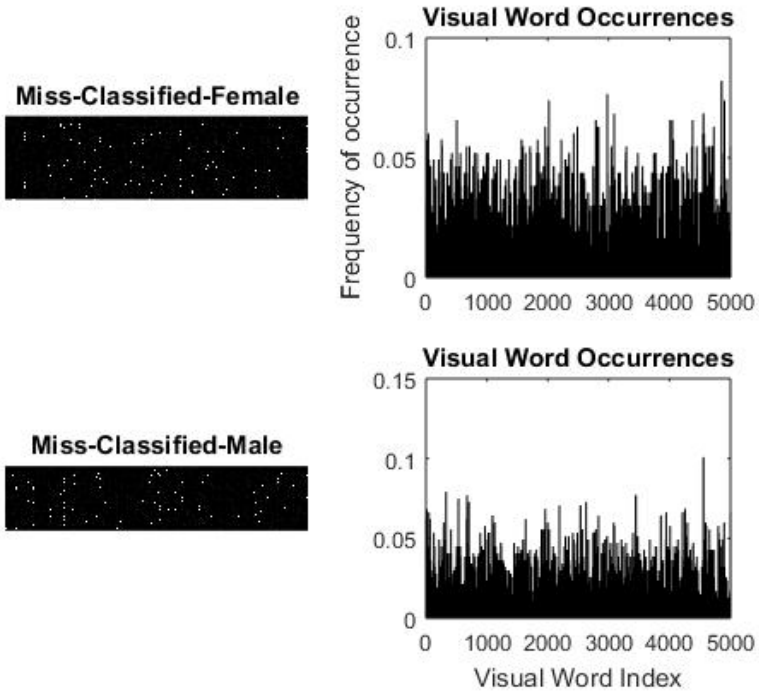

Fig. 6. Histogram of misclassified features

5 Conclusion

In this paper we have presented a generic gender identification system for classifying initiated phone calls based on the gender's sound. The process used for identifying a female voice from a male one had a precision rate of 94%. Future work will be devoted for improving the method used for classification, using Basic Local Alignment Search Tools (BLAST) algorithm [13] which is often used to compare biological sequences. This technique finds regions of local similarity between two sequences. This technique associates the nucleotides or proteins sequences to the reference databases. This method could be used to improve the classification process where our method failed due to the absence of proper reference sequences.

6 Acknowledgment

I would like to thank the Ministry of Education, The University of Hail and the Department of computer science and software engineering at the College of computer science and engineering for providing all the means necessary for achieving this research project in the friendliest working environment, my thanks is also extended to all students who volunteered their time to make this project possible.

7 References

[1] Pedro Cano, Eloi Batlle, Ton Kalker, et Jaap Haitsma. A review of audio fingerprinting. J. VLSI Signal Process. Syst., 41(2) :271–284, Novembre 2005. https://doi.org/10.1007/s11265-005-4151-3

[2] Jaap Haitsma et Ton Kalker. A highly robust audio fingerprinting system. Dans ISMIR, pages 107–115, 2002.

[3] Yu Liu, Kiho Cho, Hwan Sik Yun, Jong Won Shin, et Nam Soo Kim. Dct based multiple hashing technique for robust audio fingerprinting. Dans ICASSP, pages 61–64, 2009.

[4] Jaap Haitsma and Ton Kalker. Speed-change resistant audio fingerprinting using auto-correlation. Dans ICASSP, pages 728–731, 2003. https://doi.org/10.1109/icassp.2003.1202746

[5] Avery Wang. The shazam music recognition service. Commun. ACM, 49(8) :44–48, Août 2006. https://doi.org/10.1145/1145287.1145312

[6] Shumeet Baluja et Michele Covell. Content fingerprinting using wavelets. Dans CVMP, pages 198–207, 2006.

[7] Yan Ke, Derek Hoiem, et Rahul Sukthankar. Computer vision for music identification. Dans CVPR, pages 597–604, 2005.

[8] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. CVPR, pages 511–518, 2001. https://doi.org/10.1109/cvpr.2001.990517

[9] Martin Fischler et Robert Bolles. Readings in computer vision : issues, problems, principles, and paradigms. Chapter random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography, pages 726–740. 1987

[10] Eric Allamanche, Jurgen Herre, Oliver Hellmuth, Bernhard Froba, et Markus Cremer. Audioid : Towards content-based identification ofaudiomaterial. DansAudioEngineeringSocietyConvention110, 2001.
[11] Pedro Cano, Eloi Batlle, Harald Mayer, and Helmut Neuschmied. Robust sound modeling for song detection in broadcast audio. Proc. AES 112th Int. Conv, pages 1–7, 2002.
[12] William Pearson et David Lipman. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences, 85(8) :2444–2448, Avril 1988. https://doi.org/10.1073/pnas.85.8.2444
[13] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, et David J. Lipman. Basic local alignment search tool. Journal of Molecular Biology, 215 :403–410, Mai 1990. https://doi.org/10.1016/S0022-2836(05)80360-2

8 Author

Kawther A. Al-Dhlan has received her Phd degree from International Islamic university Malaysia (IIUM), she possesess an experience of more than 21 years in the field of teaching and research, she has more than 12 publications through different journals and conferences, moreover her thesis was recommended to have patented and published, it was a unique study for using Data mining to build Hadith classifier, currently she is working as assistance professor in computer science and software engineering department, College of computer science and engineering, at University of Hail, P.O Box 7637 K.S.A (kawthar.al.dahlan@gmail.com).

Article submitted 23 January 2017. Published as resubmitted by the author 25 February 2017.