Current Development in Artificial Intelligence.
ISSN 0976-5832 Volume 5, Number 1 (2014), pp. 11-15

http://www.irphouse.com

Age Based Cataloguing of Female Speakers Using A Naive Bayes Classifier

Nachamai M

Dept. Of Computer Science,
Christ University, Bangalore, India.
nachamai.m@christuniversity.in

Abstract

The age of speaker is a vital ingredient for many applications. Applications where access is based on age group can automatically use this simple system to identify the same. Access to machine or any equipment can also be authenticated with this system. Although there exist many difficulties when working with voice, it is a rich source of abundant information within it. With the help of acoustic and prosodic correlates of speech a model is developed in this research work to catalogue the voices into three different age groups. The MIT female prosodic database with 7 different data encoded in wav file format was used for testing. A Naive Bayes classifier was employed to do the classification into the three groups respectively. The classifier has worked on seven features extracted from the raw voice files and yielded 78.57% accuracy for classification. It is quite an interesting fact that voice embeds the age of a person and a proof can be given for the same.

Keywords Speech signal, Acoustic parameters, Naive Bayes Classifier.

I. Introduction

Speech of a person is accompanied with enormous information both in direct and hidden form. For instance direct form information could reveal to the listener the quality of voice and its tone. There are a few hidden properties in speech or voice that can be identified like gender, age, etc,. This research work has primarily attempted to find the age of a person with his/her voice input. At the very outset identifying age of a person may look very simple and naive. But the fact is, it is not a trivial information or direct data that can be computed out from the voice of a person. No one can deny that voice of a person changes with age. Scientific and physiological proof exists that voice can be mapped to age of a person. It is also true that psychological factors also

change the tone of voice, but it does not disturb the acoustic correlates of voice. There are two compromises being made for this identification process first, it is a well known fact that voice of a person does not remain the same in all occasions as well at all times. Second, the quality of voice captured may not be of complete clarity.

II. Literature review

Speaker's age detection through voice is generally estimated with acoustic correlates, linguistic and utterance information. N. Minematsu et.al[1], has attempted to automatically identify speaker's age from only acoustic information. This is generally a methodology adopted to identify a speaker. The method implemented worked well for identifying elderly speakers. In paper [2] seven different systems were combined to identify speaker age and gender identification. The base models used were Gaussian mixture model, support vector machine(SVM), and SVM based on 450-dimensional utterance level features. The method has used both acoustic and prosodic information. Mohamad et. al [3], has attempted a new approach for age estimation from telephone speech patterns based on i-vector, then a support vector regression is employed to estimate the age of the speaker. The work was tested on NIST 2010 and 2008 evaluation databases. The paper by Sedaaghi M H[4] on comparison of study of gender and age classification of speech signals has clearly explained the models used and methodology applied across age and gender identification. Automatic prediction of speaker age using Classification And Regression Trees (CART) is implemented by Sussane[5]. The CART was able to identify 72% of age group correctly. Wendy Clerx has extensively attempted automatic speech recognition using Native Dutch speakers[6]. The base system adopted in the work was Sonic large vocabulary continuous speech recognition system. Florian Metze et. al[7], has compared four approaches to age and gender recognition for telephone applications. The models adopted were a parallel phone recognizer, a dynamic Bayesian network, a linear predictor, and a Gaussian mixture model. In [8] a complete and elaborate perceptual study of speaker's age is discussed. The features of lowering of breath functions, muscle relaxation, progressive tonal lowering, lowering speech rate, increase of jitter and shimmer, lowering of formant frequencies, longer vowels and stop consonants, and increased standard deviation which directly shows impact in voice of speaker is discussed elaborately. In [9], the perception, analysis and synthesis of speaker age are extensively dealt in detail.

III. Methods and Materials

The voice of a person has acoustic correlates to age, the parameters identified for mapping to age is pitch of the voice, number of syllables uttered per second, number of phonemes per second, segment duration, spectral noise level, and jitter or shimmer levels. Pitch of voice happens to be an integral part of human voice. Pitch of the voice is actually the "rate of vibration of the vocal folds". The varying vibration changes the sound of voice. Vibrations are directly proportional to pitch, as the vibration increases pitch increases. The vibrations basically depends on the length and thickness of the

vocal chords, the reverberations are due to the contractions and expanding of these vocal chords. Change in pitch is called inflection, which happens with the change in these vocal chords, and sometimes due to health reasons and situations. Syllable is a unit of pronunciation having one vowel sound with or without surrounding consonants, forming a whole word or part of a word. A syllable can be composed of a central peak of sonority ie. sound, which may be clustered or surrounded with consonants. The central segment of the syllable will be the highest peak in the uttered word. Syllable is used as a parameter as different speakers pronunciation varies based on their native language and slang. Phonemes are picked as a parameter as it helps in distinguishing one word from another, based on the distinct units of sound. Voice intensity is the power in the voice. It is calculated as power per unit area. It is synonymous with loudness or volume in voice. There is no denial on the factor that the intensity is a varying factor based on external factors and the situation in which it is measured. Segment duration is the voice segment used for experiments where a timestamp measure for the three categories for a normalized behaviour the segment durations is identified and computed as 1, 2 and 3 respectively. Presuming, as the age increases the quantity of words, and syllables may seem to be lesser than earlier, even practically. Considering a longer segment duration would be a normalized aspect in terms of voice. Phase noise is the short term random fluctuations in a phase of a voice wave plotted. In voice peaks the phase noise can be easily identified as the peak would be represented as a blunt end rather than pitch peaks. The flat pitch peaks identifies the phase noise variation in voice. Jitter/shimmer is a rough sound accompanied with the voice, the higher the pitch the jitter happens to be more. The abrupt and unwanted variations in an interval between successive pulses happen to be more in an aged voice, which is attempted to be captured as jitter in voice. Shimmer also has the same relation to pitch but the additional variation increases the loudness in the voice. When the amplitude goes up and down in the voice we term it as shimmer.

A Naive Bayes Classifier (NBC) is implemented which will classify the age into one of the three categories using the seven features. A NBC is a probability based classifier which will work on independent observations and assumptions [10]. An independent feature model was used as the features in speech may or may not be related. The model uses maximum likelihood method rather than a supervised environment. The reason for the choicewaseven with lesser amount of training data the classifier would perform well.

IV. Experiments conducted

The experiments were conducted on the MIT university dataset which consisted of only female voices [11]. The database had different parts of speech like English vowels, English consonants, sentences, rainbow passage, spontaneous speech and prosody sentences all in the format of .wav files. There were totally 34 wav files in the dataset spoken by different female speakers of different age groups. The wav files vary in time duration starting from 30 seconds to 2.51 minutes. The experiments were conducted with the aim of identifying 3 age groups, 10-yrs to 20 yrs, 21 yrs to 45 yrs

and 46 yrs to 65 yrs. The age groups were categorized as group 1 comprising of children and teenagers, group 2 middle aged and group 3 old aged. The categorization is believed to have a correlation with the voice change with the physiological age of the larynx and voice box. The parameters shortlisted for arriving into these age groups are the basic acoustic parameters that can be derived from any speech segment or file.

Table1. Features extracted and its inference

Speaker age	10-20 yrs	21-45 yrs	46-65 yrs
Parameters			
F0/pitch	Increase	Increase	Decrease
no. of syllables/sec	4	3	1
no. of phonemes/sec	4	3	2
Intensity	more	average	Less
segment duration	1	2	3
phase noise	less	Average	More
jitter/shimmer	less	Average	Increase/more

The basic count of phonemes and syllables decrease as age increases, they happen to be inversely proportional to the age of the speaker. There seems to be a steady increase in pitch at the early categories of the age and a steady decrease at the end category, the voice box drops fatally down after 50. The segment duration increases for the aged group category as the speech becomes slow, in-turn paves way for additional spectral noise and jitter inclusions in voice.

V. Results

The feature parameter extracted was given to the NBC to group the acoustic correlates into one of the groups. The dataset had 8 wav files in group 1 category, 14 in group 2 and 12 in group 3. The proposed method performance is depicted in the confusion matrix.

Table2. Confusion Matrix

Confusion matrix	Group 1	Group 2	Group 3
Group 1	4	4	0
Group 2	0	12	0
Group 3	0	2	12

The true positive rate for the group 1, 2 and 3 are 50%, 100% and 85.71% respectively. There is more overlap in the group 2 category probably because the parameters chosen were not able to clearly distinguish the group 2 so ultimately an increased false positive rate appears in this category. The system performance value is 78.57% which is quite satisfactory.

VI. Conclusion and future work

The system is been tested on real time recordings, and the results seem very satisfactory. The work has focused on three categories of age group which can further be bifurcated for better applicability. The dataset is an attempt on female voices which can be attempted for male voices. The model is a direct approach that works on raw speech data captured, and an NBC is used for the classification which can be substituted for better outputs.

References

[1] Minematsu, Nobuaki, Mariko Sekiguchi, and Keikichi Hirose. "Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers." In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I-137. IEEE, 2002.

[2] Li, Ming, Kyu J. Han, and Shrikanth Narayanan. "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion."*Computer Speech & Language* 27, no. 1 (2013): 151-167.

[3] Bahari, Mohamad Hasan, M. L. McLaren, and D. A. van Leeuwen. "Age estimation from telephone speech using i-vectors." (2012).

[4] Sedaaghi, M. H. "A comparative study of gender and age classification in speech signals." *Iranian Journal of Electrical & Electronic Engineering* 5, no. 1 (2009): 1-12.

[5] Schötz, Susanne. "Automatic prediction of speaker age using CART." *Working Papers, Lund University, Dept. of Linguistics and Phonetics* 51 (2005).

[6] Clerx, Wendy. "Automatic Speech Recognition of Native Dutch Speakers with Different Age and Gender." *Delft University of Technology* (2007).

[7] Metze, Florian, Jitendra Ajmera, Roman Englert, Udo Bub, Felix Burkhardt, Joachim Stegmann, C. Muller et al. "Comparison of four approaches to age and gender recognition for telephone applications." In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV-1089. IEEE, 2007.

[8] Pettorino, Massimo, and Antonella Giannini. "The Speaker's Age: A Perceptual Study."

[9] Schötz, Susanne. *Perception, analysis and synthesis of speaker age*. Vol. 47. Lund University, 2006.

[10] Baesens, Bart, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. "Benchmarking state-of-the-art classification algorithms for credit scoring." *Journal of the Operational Research Society* 54, no. 6 (2003): 627-635.

[11] http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-542j-laboratory-on-the-physiology-acoustics-and-perception-of-speech-fall-2005/lab-database/