Research Journal of Biological Sciences 11 (2): 67-74, 2016
ISSN: 1815-8846

Estimating Age and Gender for Speaker through Distorted Voices Based on Fused Model

Romany F. Mansour and Abdul Samad A. Marghilani
Department of Computer Science, Faculty of Science,
Northern Border University, Arar, Saudi Arabia

Abstract: Gender and age evaluation for speech usages is very significant. One among the uses is that it can enhance person-machine communication, e.g., the announcements can be focused founded on the gender and the age of the individual on the receiver. It can assist in recognizing illegal case suspects or reduce the amount of suspects. In this essay, the estimation of gender and age was carried out using an education algorithm, including contrasting the behavior of the mechanism. Moreover, the dataset incorporated real-life experiences, such that the mechanism is compliant to real world uses. Shifted Delta Cepstral (SDC) is mined by means of Mel Frequency Cepstral Coefficients (MFCC) and the benefit of SDC is that it is further strong under loud information. From the trials, an amalgamation of MFCC and pitch was employed to get even superior acknowledgment rates.

Key words: Gender recognition, support vector machines, speech, DCT, pitch

INTRODUCTION

The focal point of this paper is the tone among the performance features in biometric gender and age identifiers for language uses with numerous realistic uses such as person-computer communication or data recovery. Furthermore it can enhance the simplicity of the mechanisms and can be supportive in orator identification and observation mechanisms.

The recognition of age and sex mechanism comprises two sections. The initial section concerns preprocessing and prospect removal. The second section includes categorization. In the initial section, language is pre-processed by means of Digital Signal Processing (DSP) systems and then several helpful characteristics like pitch and MFCC data are removed from the language.

After that these language characteristics are put into apparatus training categorizer and the mechanism is qualified with these language vectors. Within the categorization area, a choice is arrived contrasting these characteristic vectors and discovering the finest equivalent by means of SVM.

The inspiration of this work is to put into practice a strong age and sex identification mechanism for speech uses still reliable under loud circumstances. This analysis will employ the authority of the high-tech mechanism education algorithms for model recognition. The key objective in this study is to enhance sex and age recognition for language uses such that within real-life circumstances, the mechanism will offer superior recognition principles. The first step reviews the speech characteristics united with Support Vector Machines (SVM) to generate such a recognition mechanism. The next step devises a literature-independent sex and age identification system.

The third step uses strong demonstration of language indication for gender and age recognition (future assortment) such that characteristic vectors offer the finest performance. The fourth step incorporates gender and age recognition system into language identification system. The fifth stage assesses the age and gender identification mechanism in genuine world events and determines the system functionality under noisy circumstances.

There are some grounds that demonstrate that automatic gender and age recognition is not a simple duty. One basic obstacle in the initial stage is that every individuals language feature is exclusive and that makes categorization an encumbering task to begin with.

Moreover, the sound aspect presents yet another barrier. Sound can be something extra compared to the orators accent. The other part of this research is planned in the following manner. In the subsequent areas, every section of the gender and age identification mechanism is defined in facts.

Corresponding Author: Romany F. Mansour, Department of Computer Science, Faculty of Science, Northern Border University, Arar, Saudi Arabia

Part 2 entails several digital indication techniques for pre-processing. Segment 3 includes preprocessing approaches like voice activity detector and phantom calculation. Part 4 clarifies the prospect removal factors and the future removal approaches for audio-founded and pitch-founded samples. Part 5 deals with system plan and execution of the system and gives a variety of test outcomes as well with the ultimate step showing the conclusions and need for further research.

Literature review: The behavior of the gender and age recognition system relies on the speech characteristics employed. It is the subject of an investigation since the last years of the previous decade. Sedaaghi (2009) employed numerous categorizers for gender and age recognition mechanisms. These categorizers comprise possible NN, KNN, SVM and GMM.

He employed two diverse files for this task. The initial file he employed is DES catalog and the additional one is ELSDSR (Feng, 2004). PNN and SVM did finest in stipulations of sex recognition and identification of age, respectively. In 2010, Tobias Bocklet and his associates expanded the sex and age identification mechanism which was based on numerous other mechanisms and their amalgamation.

They employed spectral characteristics, prosodic attributes and glottal qualities in their function. Their finest mechanism employed GMM-UBM as categorization of sex and age. The categorization precision for this mechanism was 42.4%. Schotz (2007) reviewed audio communication characteristics like communication pace, noise force (SPL), basic incidence (F0) and maturing of language generation system in age recognition.

Dobry *et al.* (2011) suggested a novel communication dimension decrease technique called WPPCA. They tried this technique on two different occasions. The primary goal was age set categorization, followed by exact age assessment. They employed an SVM through an RBF kernel.

The presentation they viewed was superior having this size decrease method and because of the reduced speech characteristic vector dimension, the learning of SVM was a lot quicker and less prone to over-fitting.

Metze *et al.* (2007) examined four diverse approaches for gender and age identification for telephone uses, creating a contrast among persons and their mechanism on similar information sets. These methods were a corresponding phone identifier, a vibrant Bayesian system uniting numerous prosodic characteristics, linear forecast study method and finally GMM founded on MFCC for division of gender and age.

They have found that the initial means, corresponding telephone identifier was as superior as persons in conditions of recognition. However, the accuracy decreased on brief sounds. Bahari (2012) and his associate did an additional research on orator age assessment, employing WSNMF and HMM. They used Least Squares Support Vector Repressor (LS-SVR) in the capacity of a categorizer.

MATERIALS AND METHODS

Communication processing

Background: Communication signals formed by individuals are naturally analog. As a result, for computers to synthesize language data, the language should be transformed from analog to digital. Moreover, communication signs can be symbolized in occasion area or within incidence sphere whereas symbolizing language indications in occasion realm, occasion is on the x-axis and amplitude on the y-axis.

That does not inform greatly concerning the incidence speech content. Therefore, an enhanced demonstration would be the occurrence realm speech demonstration. In this realm, there is frequency on the x-axis and there is magnitude on the y-axis in dB. In addition, communication signs in occasion realm can be categorized into 3 classes: spoken, unspoken and quiet communication as shown in Fig. 1.

Spoken noises are naturally episodic and have advanced power compared to noise-like and periodic unspoken sounds. The stillness is once there is no language and might contain power height connected to the environment sound.

Speech signal features: As persons talk, the language signal generated through the verbal tract is a signal of analogue in nature. As previously stated, the information

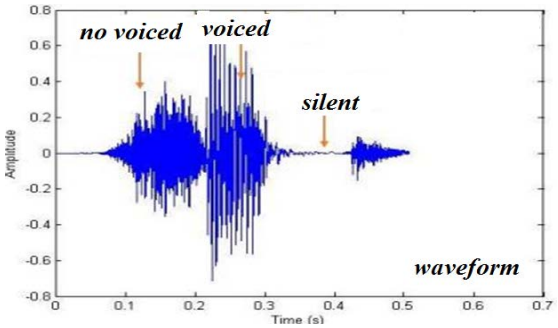

Fig. 1: Time field categorization of the language expression >Sit

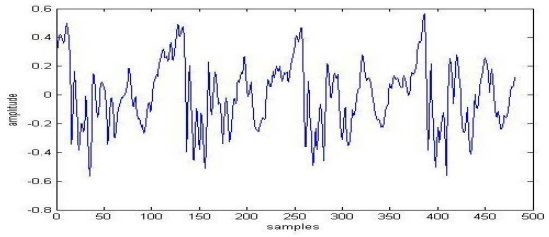

Fig. 2: One border (30 ms) of /I/ noise from the communication utterance 'Sit'

on the processors is digitally hoarded. Therefore when operating and synthesizing speech information on a processor, the primary thing is that the language data needs to be transformed into digital sign. Moreover, the communication information should be tested at extreme speeds so, as not to misplace significant data of communication. Nyquist Shannon within his modeling theorem, states that the digital testing rate should be at least two times larger compared to the uppermost rate within the analog sign.

Communication outlining and windowing: Communication sign is an occasion-changeable fixed, indication, eventually shifting. Therefore, for language to be synthesized, it should be separated into non-fixed edges. The universal dimension of language frames differs, ranging from 10-40 ms where language is considered to be non-varying. Figure 2 shows one outline of the vowel /I/ from the language expression Sit.

The communication indication is immediately scored into structures, the subsequent stage is within numerous situations are windowing. Fundamentally, the language edge is reproduced through a casement purpose. The main essential windowing purpose is the rectangular pane. Once the sky-lighting is used on the edge, not one of the standards of the edge shifts. That does not form a fine work in several situations since there are terminations on the ends of communication indications. It makes it difficult to examine those signs. Thus, to decrease this frame outcome, rather than employing a fundamental rectangular pane, several softer windows like Hamming Window are favored. Going nearer to the frames, the merit of the window draws nearer to 0 and in the center, the pane significance is 1.

While reviewing long communication signs, the language panes appear one subsequent to the other and if softer window works like Hamming are employed, it is necessary to overlie the windows to offer signal permanence. As can be seen from Fig. 3, the center

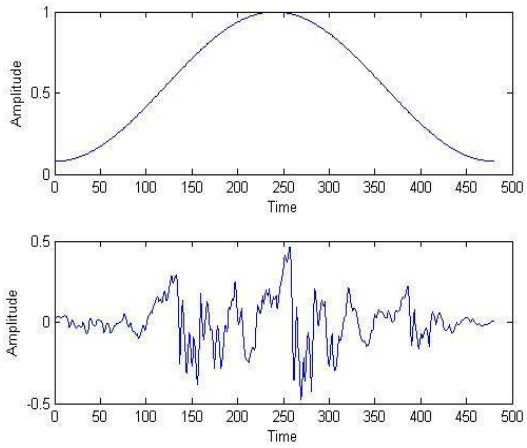

Fig. 3: On top: hamming casement, on base: hamming sky-lighted language edge of /I/ from the language expression 'Sit'

section of the indication is maintained while the indications on the edges are pointed. Therefore, to maintain indication permanence, windowed communication edges should be overlaid.

Distinct cosine change: Distinct cosine change is an indicator change that is comparable to distinct fourier convert. It shifts the indication from time realm into rate sphere in conditions of summation of cosine purpose with diverse incidences. However, it has several other benefits compared to DFT such as being more organized. It can estimate the communication indicator by means of lesser coefficients. It is employed in solidity of information, for example communication information, picture information etc. The design is that the majority of the communication incidences happen in small rates, thus DCT maintains the low language incidences that make highest of the language and eradicate extreme rates. There are numerous diverse disparities of DCT. The main normally employed DCT can be estimated as follows:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], k = 0,\ldots\ldots\ldots,N-1$$

where, x_k is the kth coefficient of DCT. Figure 4 shows the distinct cosine alteration of the language expression Sit.

Signals to sound ratio: Indication to sound proportion is percentage of the actual sign to the sound generated through the tracing gadget. Every tracing gadget has various type of sound in it. To calculate indication to

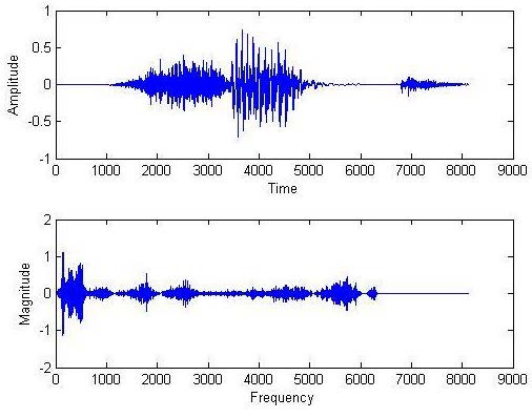

Fig. 4: Time realm demonstration of 'Sit' on base: DCT used to the communication signal

sound proportion, there are two stages in the procedure. The initial procedure is devoid of using any participation indication to the contribution, i.e., gauging of the output. Within the subsequent section when a language wave is used to the participation, the productivity of the machine is traced. This value is alienated through the significance in the initial section. In numerous situations, the sound of the machine can be eliminated. However, if the key in noise file is extremely feeble, even a minute sound from the machine can result in deformation. The indication presentation can be interpreted by viewing the SNR. The bigger the sign to sound ratio is the superior the excellence of the language. Moreover, indication to sound proportion is calculated in dB that makes it simpler to handle large amounts. Indication to sound proportion (SNR) can be calculated by means of this formula below:

$$SNR = \frac{\rho_{signal}}{P_{noise}}$$

Where:

ρ_{signal} = The mean authority of the indication
P_{noise} = The mean influence or the sound

This can be stated in logarithmic conditions employing dB:

$$SNR_{dB} = 10\log_{10}\left[\frac{\rho_{signal}}{P_{noise}}\right]$$

Figure 5 demonstrates an instance of a sine signal by means of sound inserted to it, where Pn/Ps = 0.1.

Noise removal methods: Spectral Sound is, until recently, among the main broadly applied language sound extraction approaches. It supposes that language has

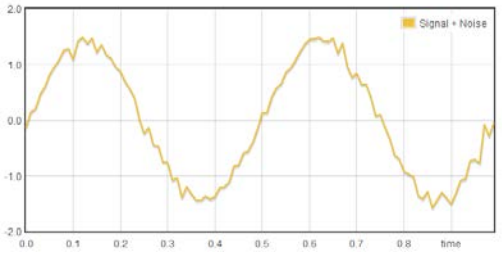

Fig. 5: Instance of sinusoidal with waves to sound proportion Pn/Ps = 0.1

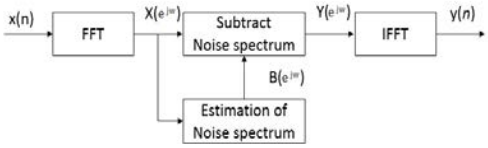

Fig. 6: Spectral calculation approach

actual speech information and various backdrop sounds in it. This method functions in the incidence realm and it presumes that the loud participation communication can be articulated in conditions of language range and the backdrop sound continuum.

Figure 6 depicts the map of spectral sound calculation technique. The sound continuum is approximated from the language models where merely a sound is available and then it is deducted from the innovative loud language range to obtain the clear communication range.

In Fig. 6, x(n) is the distinct time loud speech succession, whereas in x(n) = s(n)+b(n) where s(n) is the clear wave and b(n) is the backdrop sound where $X(e^{jw})$ is the FFT of x(n), $B(e^{jw})$ which is the evaluation of the sound continuum. So, $Y(e^{jw}) = X(e^{jw})+B(e^{jw})$ and lastly y(n) is the distinct time clear language series by using inverse fourier transform.

RESULTS AND DISCUSSION

Voice activity detectors: Tone Activity Detector is among the mainly broadly applied pre-processing approaches within language uses (Sohn *et al*., 1999). It fundamentally attempts to decide which language edges have language and which do not contain speech edges. When a person talks, he or she cannot talk without pausing, thus there will be numerous gaps and perhaps some faltering etc. in the speech. Therefore, to remove some helpful speech characteristics from the communication, those non-language noises or stillness must be removed from the language models. Several additional uses of VAD can be observed in acoustic consultations, echo annulment, language identification and language programming and hands open telephony (as cited voice activity detection).

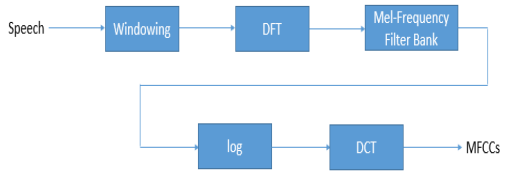

Fig. 7: Stages of MFCC calculation

In general, right following communication is digitized; VAD is used to the communication indication.

Mel incidence Cepstral Coefficients (MFCC): As stated previously, MFCC is one of the mainly extensively applied language removal aspect in language uses. Language is created using a person=s choral tract as well as tongue and lips. Therefore, to identify what has been stated, the form of the choral sieve should be modeled. In this regard, MFCC attempts to sample this verbal tract sieve in short occasion authority continuum. MFCC was 1st founded by Davis and Mermelstein (1980).

Since then, it has been the high-tech audio communication aspect. Prior to the creation of MFCC, some extra removal techniques were used like Linear Forecast Coefficients and Linear Prediction Cepstral Coefficients (LPCC). Figure 7 illustrates the MFCC calculation.

The initial stage in calculating MFCC is the sky-lighting. Language is cased into 20-40 ms edges. In case the communication period is less there may not be sufficient samples for a dependable calculation or else if the communication period is above 40 ms then, there will be numerous shifts within the indication. The second stage is where the control continuum of every edge is computed.

At the 3rd stage, the time gram is established and entails numerous data concerning speech. Essentially, if the rate rises, cochlea cannot function as well at discriminating the incidences. At small rates, sieve bank is thinner and at extreme incidences sieve bank turns out to be broader. Thus, mel-frequency sieve reservoir is applied to decide the power heights of incidence areas. In the subsequent stage, the log is removed from the sieve bank powers.

The rationale for this is that an individuals ear does not contain a linear range in terms of hearing. The last stage involves getting the DCT of the log sieve bank powers. The ground for executing that is to connect the overlaid sieve reservoir powers for improved categorization. In the subsequent step whereas changing from rate to Mel balance this method is applied Fig. 8 plot of mel frequency range:

$$(f)=1125\ln\left(1+\frac{f}{700}\right)$$

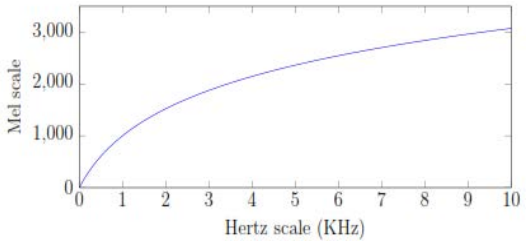

Fig. 8: Plot of mel frequency scale

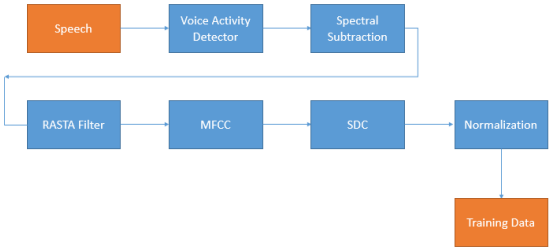

Fig. 9: Block drawing for sdc characteristic removal algorithm for the era and sex identification mechanism

Features extraction: The two aspects were applied for gender and age recognition in this scheme. The initial characteristic applied was SDC. As stated formerly, SDC is a derivative from MFCC. Subsequent to using various preprocessing approaches, MFCC was computed through the 12 coefficients with every edge dimension being 30ms. Subsequent to MFCC computation, a RASTA sieve was used to the wave to get rid of the canal sound.

Afterward, SDC was computed employing the documentation having the N-d-p-k factors set to 7-1-3-7. After this stage, the average value of SDC was regularized to zero and the average divergence was standardized to one. The block drawing of attaining SDC aspects can be viewed in Fig. 9.

For attaining the other communication characteristic pitch, the records given by Sun were applied. The algorithm employs sub choral to choral proportion to obtain the pitch data. In addition, the incidence variety was positioned amid 100 and 300 Hz for extra dependable pitch assessment.

Every edge dimension was programmed to 25 ms. After attaining the pitch data for the entire edges in the learning group, the average value for every preparation set was assumed as the terrain of the learning instance.

Tests and outcome: This paragraph contains all the findings and the presentations of the trials. As stated previously, an age and sex identification mechanism can be established using several different means. This

mechanism can be established by means of diverse speech characteristics as well as combining several speech characteristics.

Moreover, all the trials that are conducted here employed a locked-set which implies that the learning and trying information was from a similar basis. The objective of doing diverse tests is to contrast various communication characteristics in studying sex and age identification mechanism. As stated previously, dissimilar speech aspects were applied to check the mechanism. These aspects are the pitch rate, MFCC and lastly SDC. In addition, improved means of uniting MFCC and pitch were tested. SVM categorizer having nonlinear RBF root was coached regarding pitch data.

The categorizers choose a nonlinear porch to divide between every label. For the additional characteristics SDC and MFCC, once more a nonlinear SVM having RBF kernel was prepared. SVM was chosen as a categorizer due to its authority and easiness of application. Moreover, for all the evaluations, the equivalent preparation experiment was applied, thus a strong assessment on the presentation could be done.

Ageand sex identification is a multi-set categorization task. After employing the earlier pre-processing and SDC characteristic removal stages, an SVM was educated for mutual sexes and age sets.

Nonlinear RBF root was employed during this experiment. The file had 4 tags. These stickers integrated a young grown-up male and a young mature female ranging from 20-40 years as well as mid-aged male and female ranging from 40-65 years. For SVM learning, the algorithm defined in this work was employed (Chang and Lin, 2011). The factor of SVM was initially chosen by hand and at the next occasion they were chosen with traverse legalization after learning on an evenhanded sub-group.

Pitch based models: Pitch-founded pattern employed merely pitch data of the orator to identify his or her age and sex. At first, preparation information was loaded into Matlab R2010a. After that, the pre-processing algorithms like VAD and ethereal calculation were applied to eliminate the stillness and the backdrop sound from the teaching.

Furthermore, pitch removal algorithm that employs choral to sub-choral proportion was implemented. The basic incidence standards vectors have the intended primary rates for the entire edges. To obtain the real pitch of the preparation instance, the average significance was computed in the conclusion. Consequently, for the categorization intention, a non-linear SVM was educated. Within the trying section, the similar pre-processing, sound improvement and removal methods were used on the experiment information and this value was conveyed into SVM for categorization.

A variety of experiments were conducted to better understand the pitch and age and sex identification connection. For this experiment, a non-linear SVM having RBF core was used for categorization of age and sex. The preparation group comprised sounds from speakers of each age and sex set. The whole preparation occasion was around 6 min in total. The pattern was tested for the experiment class. The experiment outcome can be seen in Table 1.

From the chart, it can be assumed that categorization proportion for group 4 is less compared to the additional groups. Moreover, generally set accuracy is 50%. Moreover, it is likely that pitch data was overlaid with the additional groups. Therefore, it was faulty to assume that employing one speaker from every sticker to identify age and sex can have positive results.

Figure 10 shows the group signs and the characteristics. As stated above, the preparation set comprised one speaker of each age group. Thus, in the preparation group, it can be observed that there are four impartial classes. However, when it gets to the experiment group, the initial class has additional instances than the additional categories as can be observed from the form. Moreover, the figure also shows the characteristics of pitch data.

Table 1: Pitch founded replica outcomes for one orator for every sex and age set

Groups	Value (%)
1	40
2	100
3	50
4	10
General group precision	50

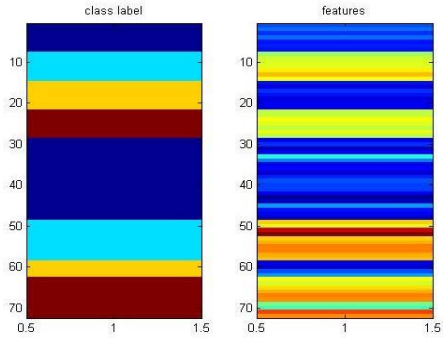

Fig. 10: Set tags and characteristics for pitch founded pattern for one speaker for every sex and age set

Figure 11 presents the categorization outcomes. The empty dots symbolize the models from the preparation set. Packed dots stand for the information from the experiment set. Moreover, the various colors symbolize diverse groups allocated by SVM categorizer. In addition, frame color represents the exact sticker of the illustration. The dimension of the dots symbolizes the assurance height of that instance if it is better, the self-assurance point is superior.

Pitch and MFCC fused model: It is a merged sample of pitch rate and MFCC. Preparation of the pattern was clarified at the beginning of this study. Experiment outcomes can be seen in Table 2.

To combine two categorizers, the likelihood matrix from terrain-founded form was over-modeled to go with the dimension of the likelihood matrix from the MFCC founded replica. Then, the likelihood standards of both categorizers were standardized to attain more precise outcomes. Two diverse coefficients were used for the categorizers. According to the examination, pitch-founded sample was offered a coefficient of 0.6 and MFCC founded sample were offered a coefficient of 0.4. Lastly, a weighted amount was used of the likelihood matrices to get the last chances.

From Fig. 11, it can be observed that the general precision is simply 0.8% beneath the pitch-founded form. This does not indicate a large discrepancy. The excellent obsession concerning merging the scales of 2 categorizers

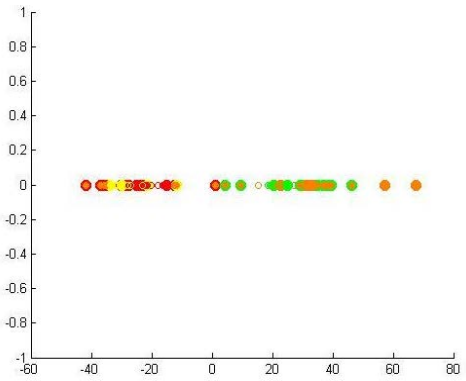

Fig. 11: Categorization outcomes for pitch founded era and sex identification

Table 2: Outcomes from pitch and MFCC merged sample

Groups	Value (%)
1	91.8
2	67.6
3	24.1
4	73.0
General group exactness	64.2

is that group 3 precision increased to 74.15% which implies that it has extra likelihood to properly notice certain ages and sexes. Group 3 precision was low for almost the entire duration of the experiment. The motive that class 3 accuracy is low is due to small number of preparation models in dimension. Moreover, it can be distinguished that factor choice is actually significant in SVM preparation with RBF core.

Because of the occasion and material boundaries, experiments were not carried out on a dissimilar or larger data group. However, from the merged form it can be observed that the pattern functionality is maintained.

CONCLUSION

The objective of this work was sex and age identification for speech uses. To construct that type of mechanism, all essential stages were clarified in detail. These stages incorporated pre-sign processing methods, communication characteristic removal methods and lastly categorization algorithm. Three diverse models were projected for this task. Within the initial sample, the pitch data was applied to attempt to identify sex and age of the speaker. Within the second sample, SDC and MFCC were applied to determine language characteristics which completed the categorization stage. In the final sample, a merged mechanism was planned that combined field and MFCC.

Different experiments were executed with diverse information dimensions and various categorization elements to calculate the presentation of the mechanism. Because of the occasion and material restrictions, the experiments were conducted on a comparatively small data group.

According to the experiment outcomes, it was observed that MFCC did great as a solitary language aspect. The motive for this was made clear in the previous parts. Moreover, the best identification value was attained by means of the pitch and MFCC merged sample. This merged sample gave a precision of 64.20% when tried on ELSDSR (Feng, 2004) file and with sensible C and gamma factors.

RECOMMENDATIONS

Future research should focus on increasing the performance of the mechanism, by introducing and executing some extra steps. Furthermore, the training information was barely noticeable in dimension.

Therefore, with the preparation information size much larger, the mechanism behavior can become improved. The mechanism can be used on higher processors so as to abridge the preparation occasion. Several loud data groups can be applied and particularly SDC-founded sample experiments can be seen. Pitch, SDC and MFCC were also applied in this experiment. Therefore, various additional language characteristics can be put in into the mechanism that can ultimately augment the system presentation.

REFERENCES

Bahari, M.H., 2012. Speaker age estimation using hidden markov model weight supervectors. Proceedings of the 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), July 2-5, 2012, IEEE, Montreal, Canada, ISBN: 978-1-4673-0381-1, pp: 517-521.

Chang, C.C. and C.J. Lin, 2011. Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol., Vol. 2, No. 3. 10.1145/1961189.1961199

Davis, S. and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process., 28: 357-366.

Dobry, G., R.M. Hecht, M. Avigal and Y. Zigel, 2011. Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. IEEE. Trans. Audio, Speech, Lang. Process., 19: 1975-1985.

Feng, L., 2004. Speaker Recognition, Informatics and Mathematical Modelling. Technical University of Denmark, Lyngby, Denmark,.

Metze, F., J. Ajmera, R. Englert, U. Bub and F. Burkhardt et al., 2007. Comparison of four approaches to age and gender recognition for telephone applications. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, April 15-20, 2007, IEEE, Honolulu, Hawaii, ISBN: 1-4244-0727-3, pp: IV-1089.

Schotz, S., 2007. Acoustic Analysis of Adult Speaker Age. In: Speaker Classification I. Muller, C. (Ed.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-540-74186-2, pp: 88-107.

Sedaaghi, M.H., 2009. A comparative study of gender and age classification in speech signals. Iran. J. Elect. Elect. Eng., 5: 1-12.

Sohn, J., N.S. Kim and W. Sung, 1999. A statistical model-based voice activity detection. IEEE. Signal Process. Lett., 6: 1-3.