Gender Identification via Voice Processing

Shivangee Kushwah, Shantanu Singh, Kshitij Vats, Varsha Nemade

Abstract: ***Recognizing the sexual orientation of a person via speech has an assortment of utilizations going via speech investigation to customizing machine and human collaborations.***

Speech Recognition Technology can be installed in different continuous applications so as to expand the human-PC association. From apply autonomy to social insurance and aviation, from intelligent voice reaction frameworks to versatile communication and telematics, speech acknowledgment innovation have upgraded the human-machine connection. Sexual orientation acknowledgment is a critical part for the application implanting discourse acknowledgment as it lessens the computational intricacy for the further preparing in these applications. This paper involves extracting the constituent parts such as frequencies and interquartile ranges which are in light of a legitimate concern for distinguishing the speaker's sexual orientation with as meager discourse as could reasonably be expected.

Index Terms: *Indexing-Optimization-Big Data-Artificial Intelligence*

I. INTRODUCTION

Speech is the most helpful mode of correspondence among the people. Discourse signals shift exceedingly alongside time and are the most arbitrary signs. Inquires about have been done so as to have such simple collaboration among people and machine [2]. As we realize that innovation is developing at a fast rate which incorporates upgrades in advances like that of huge information, development of different new calculations, machine learning and so forth. Alongside these, the field of discourse acknowledgment has likewise created.

By Learning about gender, it can be utilized for standardization of highlights of speech, which has appeared to diminish word blunder rate [1] in discourse acknowledgment. Sexual orientation recognizable proof can enhance the expectation of other speaker attributes, for example, feeling and age, either by together displaying gender with age (or feeling) or in a pipelined way [15]. Speaker confirmation frameworks likewise certainly or unequivocally use sexual orientation data.

In general, speaker's identification for sexual orientation is imperative for progressively regular and customized exchange frameworks. Past work has analyzed the different contrasts among male and female discourse. These distinctions incorporate physiological (for example length of vocal tract), phonetics, and quality of voice contrasts. According to the investigations of the human view of pitch area, total basic recurrence (f0) has been noticed to be as the most important data to choose between both related pitch levels and speaker's sexual orientation [1]. A common place way to deal with recognizing sexual orientation (or other speaker attributes) is to process outline measurements of the discourse highlights (pitch or unearthly, for example, mean, highest, lowest, over an explicit time range and to utilize attained insights as highlights in sex characterization theorems. In this paper, we have displayed the connection in sex and both prosodic and spectral capabilities, where each list of capabilities are considered exclusively just as together. We likewise present and examine a novel methodology that utilizes the whole element direction for gender identification. This methodology has the advantage of being low idleness with regards to spilling discourse input, and possibly give a dedicated portrayal of the contribution without the misfortune brought about in registering rundown insights

II. EXPERIMENT ON EVALUATING GENDER

The Data samples that were put to use is the "How May I Help You" dataset, or one can all it HMIHY, it is a corpus/dataset used in the automatic gender detection experiment. This corpus will help to experiment on the sample and then compare it with previously published results in order to inspect the accuracy of the said approach taken. HMIHY is a corpus which was generated from an NL (Natural Language) dialogue system, a human-computer spoken system which has been developed at AT&T Labs. This enables automated agents to reply to someone on a phone call, and these telephonic conversations are then summarized into an easily accessible speech data. Although, the data is not devoid of any kind of noise and background interferences.

The aforementioned corpus has data from around 1654 different speakers which include up to 5002 utterances and each utterance has an average duration of about 5 to 6 seconds. In the experiment, since gender labels were required to test, 4520 utterances have been used and the gender labels were applied by two humans with utmost accuracy and no point of error in any case of application of gender labels, so as to subject the experiment to a near perfect data set. To begin experimentation, the test sets were divided into 3 individual test in the ratio of 8:1:1.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author

Shivangee Kushwah*, Scholar, Department of Computer Science Engineering, Narsee Monjee Institute of Management Studies.

Shantanu Singh, Scholar, Department of Computer Science Engineering, Narsee Monjee Institute of Management Studies.

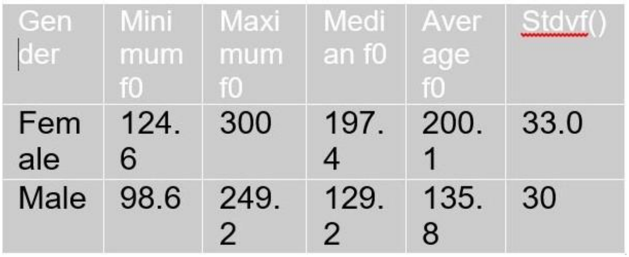
Kshitij Vats, Scholar, Department of Computer Science Engineering, Narsee Monjee Institute of Management Studies.

Varsha Nemade, Assistant Professor, Department of Computer Science Engineering, Narsee Monjee Institute of Management Studies.

That is, 80% of data was used for training the kernel, 10% for testing the kernel after it has been trained and the rest 10% was used for development purposes. An important detail to note is that these divisions or partitions in the dataset were created the percentage of speakers, each split had around 36% of female speech data to 64% speech data. This may or may not lead to some bias in the given kernel used for experimentation. The baseline for the accuracy of identification of given gender from the data set in the corpus has been at around 64%.

The table below shows the average f0 statistics for speakers of both genders in the given dataset training.

Table 1: Mean f0 statistics by male and female speakers



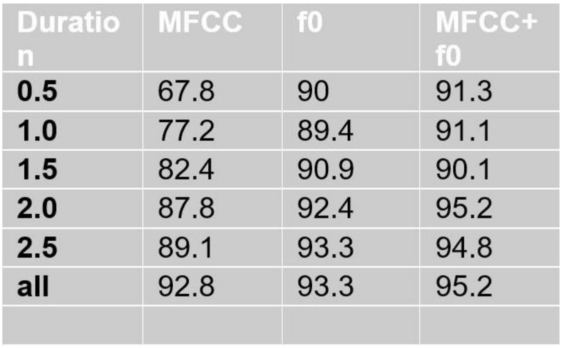
Gender	Minimum f0	Maximum f0	Median f0	Average f0	Stdvf()
Female	124.6	300	197.4	200.1	33.0
Male	98.6	249.2	129.2	135.8	30

Classification based on Rules Inferring from the statistics in the above table, one can see that there has been a substantial and large difference in f0 ranges among the female and male speakers. To be particular, one can observe that there is a large gap between females and males in the different values of the mean-f0. Using this contrast as way to understand the relationship, one can develop a very simple rule which could act as one of our main baseline classifier, i.e. if the mean f0 of the speaker is closer to the average pitch of males than that of the average pitch of females, the speaker will be labeled as male. If not so, the one speaking would be labeled as of gender female. This simple approach has been applied to the test dataset entirely, and an accuracy of around 85%+ was achieved through this method. This rule depends on knowing the relationship between the two genders and the mean frequency f0, which helps one to develop a very simple rule in order to use one's speech to detect their gender. The rest of this paper deals with attempting at improving the gender classification accuracy by not only above the majority-class baseline which is 64% but to push it also above the mean f0 rule-based baseline of 87% with the adoption different AI and machine learning techniques in order to complete the task at hand, i.e. gender detection.

III. MFCC STATISTICS AND F0 STATISTICS

Generally, properties we use to detect the gender and other similar types of efforts are present in the structure of a summary statistics of spectral and prosodic properties, especiallymade in where discriminative classification for approaching the said problem are needed. In this experimental set-ups, 4 types of classifiers have been used, they are random forest, linear and logistic regression and lastly AdaBoost. All of these classifiers were applied on the various frequency f0 summary statistics and the MFCC summary statistics as well as using a combo of both the above summary statistics.So far, the different experimentations have been carried out with different duration of spoken recorded sequences in to properly discern the relationship between how the duration of recording affects the accuracy of the classifier method used. So far, the inferred knowledge has been that logistic regression classifiers have been performing best for all the three types of data-sets mentioned above and therefore only those results have been included in this paper.

Table 2: classification results of MFCC and f0 summaries using logistic regression learning on different lengths of utterances



Duration	MFCC	f0	MFCC+ f0
0.5	67.8	90	91.3
1.0	77.2	89.4	91.1
1.5	82.4	90.9	90.1
2.0	87.8	92.4	95.2
2.5	89.1	93.3	94.8
all	92.8	93.3	95.2

We can infer from the above Table 2 that even though speech recording of as little as 2 seconds of speech is enough, the best accuracy in detecting the gender using logistic regression classifier is obtained by using the entirety of the recording available. Above results display that f0 features turn out to be more predictive of gender than any other type of features such as cepstral features. We can also infer that f0 features come out to be more capable and practical immediately at the beginning or start of a speaker's utterance; more than 91% of the accuracy is achieved only after 1 sec of the recording.

In the comparison of f0 to MFCC, one can infer that this model shows a sharp increase in the performance as the speech increases. This illustrates that features of MFCC are of more use over longer periods of speech. Also, the classifier model where both f0 and MFCC properties or features were combined are constantly more precise for all the different durations, with only 2 seconds of speech, about 95% of accuracy has been obtained which provides us admission to couple the spectral as well as prosodic features for better accuracy.

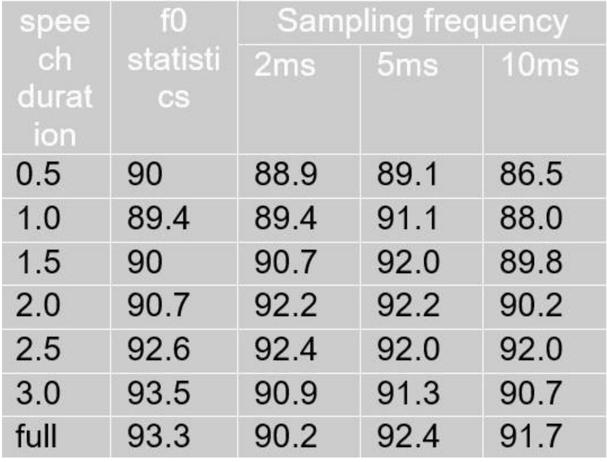
IV. F0 TRAJECTORIES

In the aforementioned method, we've made rigorous use of summary statistics based features in order to increase the accuracy of our model, instead of focusing on summary statistics one needs to explore the significance of entire raw trajectory of f0 values to detect gender. This allows us to prevent the use of extensive statistical calculation in order to reach to inferences where low latency gender detection is required. Contrasting from the f0 trajectories method, the summary statistics feature output their results in the form feature vectors which are of fixed dimensions, while in the case of f0 trajectories gives one variable dimension feature vectors of which the classifier algorithmic implementations are not allowed as input. Therefore, one needs to model the trajectories of f0 as textual value of input, that is, each token of text has to have corresponded to the value of f0 to the nearest tens place.

In this experiment, Maximum Entropy text classifier, also known as MaxEnt and LLAMA (used to compute the trigrams as features on the training sample)was used.The experimentations were conducted on three different f0 trajectories by siphoning the f0 values every 2ms, 3ms, and 10ms.

In the experiment, the LLAMA model was trained on binned f0 statistics, which would help in order to precisely compare results from the categorical properties on trajectories against the statistics. The f0 features'statistics were binned by rounding off the maximum, minimum, median and mean f0 values to the nearest tens place. [6][10]. Displayed below (table 3) we try to compare the LLAMA statistics approach with the LLAMA trajectories approach, with the 10ms sampling of the statistical approach, the statistics approach will outperforms. It outperforms the trajectory approach at almost each and every period without exceptions [10]. Yet, undeviating from our hypothesis, the trajectory approach as 5ms sampling outperforming the statistics approach of LLAMA in the following durations -> 1.5 second, 1second, 2 seconds, 2.5 seconds and 3 seconds. Although, at around 2ms sampling, the trajectory approach is better between 1.5 seconds to 2 seconds. These consequences, indicate that one could make use of trajectory approach to identify gender where applications require the fairly accurate predictions in less than or around 1 second [6].

Table 3: classification results for f0 trajectories [9].



speech duration	f0 statistics	Sampling frequency		
		2ms	5ms	10ms
0.5	90	88.9	89.1	86.5
1.0	89.4	89.4	91.1	88.0
1.5	90	90.7	92.0	89.8
2.0	90.7	92.2	92.2	90.2
2.5	92.6	92.4	92.0	92.0
3.0	93.5	90.9	91.3	90.7
full	93.3	90.2	92.4	91.7

As we can summarize for the above approach, the trajectory approach for gender identification is an innovative approach to this problem. Even though the this approach may be performing lower than the techniques of ML classifier algorithms which were used to train on the f0 statistics, it is still an unusual feat see f0 trajectory as gender identification classifier problem and that this would perform so well. Also, one might be able to improve the way it's performance and with a simple and swift method to identify gender by making use of a more sophisticated quantization method of the values of f0.

V. CONCLUSION

We present the aftereffects of some examinations recognized with altered gender identification from communication between the sexes. In our first set of examinations, we find that utilizing direct f0 rundown encounters from just 2.5 seconds of talk, we are able to accomplish indistinguishable effects from found out in [8] using MFCC and f0 consolidates in general verbalization (around 6 seconds). We present a novel going to supervise gender identification, utilizing the entire f0 course of attributes as responsibility to LLAMA, a straight out classifier. This accomplishes striking consequences outcomes mulling over that the numeric highlights are handled as n grams, and recommends that the direction technique may be vital for anchoring in reality correct gender needs with as little as one preview of communication. We don't forget MFCC and f0 highlights and find that f0 alone is more discriminative than MFCCs, but a mixture of both component streams yields the maximum simple execution. We display that using MFCC and f0 highlights from simply 2 seconds of talk, we may anchor difficult to understand consequences from [8] utilising comparative highlights in general rationalization. In our cross-lingual examinations, we find that we can set up a sexual presentation classifier on a bearably insignificant English arranging set (three.6k verbalizations, 2 seconds every) and attain a 92% exactness checking out on an expansive German corpus (18k motives, 2 seconds every), about on a general with getting ready and trying out from a close to German corpus (93.8%). At closing, our examinations with the 3-magnificence problem male, lady, young people, and result in a nearby emerge framework making use of a non-obligatory timberland show orchestrated on clear f0 and MFCC highlights. Future paintings utilizing f0 headings can employ a gradually present day binning device to all the nearly positive deal with the numeric highlights. We can enhance the 3-route classification through making use of greater youthful talk corpora for arranging. Furthermore, the scikit - examine models on this work applied the default parameters; tuning the parameters will in all likelihood result in broadened execution. Notably more considerably, the frameworks on this work can be related with distinct different paralinguistic quarter issues, for instance, age and feeling. Specifically, its miles plausible that our bearing strategy can get subtleties inside the f0 form that define estimations cannot, and might enhance execution of para semantic conspicuous verification systems.

REFERENCES

1. Sarah Ita Levitan, Taniya Mishra, Srinivas Bangalore "Automatic Identification of Gender from Speech"
2. Speech Feature Extraction for Gender Recognition I.J. Image, Graphics and Signal Processing, 2016, 9, 17-25.
3. J. Bishop, & P. Keating, "Perception of pitch location within a speaker's range: Fundamental Frequency, voice quality and speaker sex", in The Journal of the Acoustical Society of America, vol. 132-2, pp.1100-1112, 2012.
4. R. Vergin, A. Farhat, & D. O'Shaughnessy, "Robust gender dependent acoustic-phonetic modeling in continuous speech recognition based on a new automatic male/female classification", in Spoken Language, vol. 2, pp.1081-1084, 1996.
5. C. G. Henton, "Fact and fiction in the description of male and female pitch", in Language and Communication, vol.9, pp.299-311,2010.
6. E. S. Parris, & M. J. Carey, "Language independent gender identification", in Proceedings of ICASSP 1996, vol. 2, pp.685-688, 2009.
7. P. Boersma, and D. Weenink, "Praat, a system for doing phonetics by computer .", in Glot International, pp.341-345, 2011.
8. Y. Hu, D. Wu, & A. Nucci, "Pitch-based gender identification with two-stage classification.", in Security and Communications Networks, vol. 5, pp.211–225, 2012.

9. S. Wegmann, D. McAllaster, J. Orloff, & B. Peskin, "Speaker normalization on conversational telephone speech", in Proceedings of ICASSP 2015, vol. 1, pp.339-341, 2015. F. Eyben, F. Weninger, F. Gross, &
10. B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor.", in Proceedings of ACM Multimedia,Barcelona, Spain,pp.835-838, 2013.
11. I. Shafran, M. Riley, and M. Mohri, "Voice signatures", in Proc.IEEE Automatic Speech Recognition and Understanding Workshop, pp. 31–36, 2015.
12. S. Safavi, P. Jancovic, M. J. Russell, and M. J. Carey, "Identification of gender from children's speech by computers and humans.",in Proceedings of Interspeech 2013, Lyon, France, pp. 2440–2444,2013.
13. H. Harb, and L. Chen, "Gender identification using a general audio classifier", in Proceedings of ICME 2003, pp. 733–736, 2003.
14. B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "Paralinguistics in speech and language–state-of-the-art and the challenge.", in Computer Speech& Language, vol. 1, pp.46–64, 2013.
15. H. Meinedo, & I. Trancoso, "Age and gender classification using late fusion of acoustic and prosodic features", in Proceedings of Interspeech 2016, Makuhari, Japan, pp.2818-2821, 2016.

AUTHORS PROFILE

Miss. Shivangee Kushwah, Scholar, Department of Mechanical Engineering, Narsee Institute of Management Studies, Mumbai, Maharashtra, India. She has presented various papers in various conference. Her area of interest is image processing and artificial algorithms

Mr. Shantanu Singh, Scholar, Department of Mechanical Engineering, Narsee Institute of Management Studies, Mumbai, Maharashtra, India. He has presented various papers in various conference. Her area of interest is coding and artificial algorithms

Mr. KshitijVats Scholar, Department of Mechanical Engineering, Narsee Institute of Management Studies, Mumbai, Maharashtra, India. He has presented various papers in various conference. Her area of interest is coding and software development.

Prof. Varsha Nemade, Assistant Professor, Department of Mechanical Engineering, Narsee Institute of Management Studies, Mumbai, Maharashtra, India. She has presented various papers in various conference in both national and international level. She has published 10+ in national and international journal. Her area of interest is machine learning and deep learning.