Interspeech 2018
2-6 September 2018, Hyderabad

Extracting speaker's gender, accent, age and emotional state from speech

Nagendra Kumar Goel, Mousmita Sarma, Tejendra Singh Kushwah, Dharmesh Kumar Agrawal, Zikra Iqbal, Surbhi Chauhan

Go-Vivace Inc., McLean, VA, USA

(nagendra.goel, mousmita, tejendra, dharmesh, zikra.iqbal, surbhi.chauhan)@govivace.com

Abstract

We demonstrate a speaker characteristics assessment solution to extract speaker's information like gender, age, emotion, language and accent from telephone quality speech. The solution has been designed using machine learning algorithms ranging from Gaussian mixture models to deep neural networks and utilize websocket technology for real-time bidirectional interface to provide live updates in a scalable manner. The service is utilized on our demonstration web-page where user can upload or record audio file and obtain the speaker's characteristics. Such speaker characteristics information can be used as metadata in many real life applications designed for an emotionally sensitive human to machine interaction and human to human interaction.

1. INTRODUCTION

Extracting information like age, gender, language, accent and emotional state from speech has particular importance in intelligent commercial dialogue systems and smart call centers. The ability of a machine to be automatically aware of such information about the speaker can help the automated response system to give a better suited response. In case of human interactions, such as in call centers, the metadata allows for better matching between the caller and the agent who is serving the caller. For example, we can choose to connect to an agent who speaks the same language or has a similar accent. Once the connection is established, the machine may still continue to assist the human agent with suggestions for up sells or appropriate responses to a question in a better manner. In case of machine response, there is an opportunity to select the right language and the text to speech tonal qualities that better please the caller given the age, gender, accent and other related pieces of information. Even the dialogue flow could be more age and gender appropriate. A matching response accent may be better understood and appreciated. Understanding the age and gender of the speaker may help to forward the call to an agent of the same age and gender group. An emotionally intelligent machine may help to understand customer's response to the product, their psychology etc. Thus virtual reality and dialogue applications may begin to personalize themselves on the basis of these speech body language cues. In a well designed system, such an approach will improve user satisfaction.

Here, we describe such a solution to extract speaker characteristics. We report performances of individual recognition engines on some fixed dataset which we internally use to evaluate system level improvements of our algorithms. The rest of the paper is organized as follows. Section 2 describes the API and services and Section 3 describes the technology used for the recognition and classification engines. We conclude the description by reporting performance of the recognition engines in Section 4.

2. Description of the services

The information extraction solution is made available for external use via a websocket based interface. This allows easy integration into live streaming applications, including high volume call-center platforms, using a simple, widely used, real-time and bidirectional interface. A web page has been constructed to demonstrate these capabilities, and a photograph of the same is shown in Figure 1 (available at https://www.govivace.com/solutions/speaker-characteristics/). A java applet is used to record the speech and stream it live to the various characteristics recognition engines. The results are obtained within fraction of a second after the recording stops. As shown in Figure 1, the user can upload or record audio by clicking the respective buttons. The users can also listen to the uploaded or recorded audio. The **evaluate** button is available to the user to get the response from our servers, which shows the speaker characteristics on screen with respective scores. Users can also select portions of the audio and evaluate only for those portions by using the **evaluate selected audio** button.

Table 1: *Performance of Recognition engines*

Recognition engine	Classification Accuracy, MAE*
Gender	97.9
Language	83.82
Accent	35.94
Emotion	91.46
Age	5.18*

Table 2: *Duration level classification accuracy of Language and Accent Recognition engine (* not computed)*

Recognition Engine	3 sec	7 sec	10 sec	30 sec
Language	68.56	*	87.62	95.27
Accent	*	35.15	42.82	*

3. Description of Technology

Each of the components use slightly different learning algorithms. We describe the inside technology of individual recognition engines below.

The gender, language and accent identification engines use I-Vectors trained on top of a DNN acoustic models as described in [1]. Individual secondary classifiers are trained on the I-Vectors, using multinomial logistic regression. The system structure is depicted in Figure 2. The accent identification system was trained to detect 9 different English accents at present. The language identification system was trained on 13 languages including English using logistic regression in the final classification step.

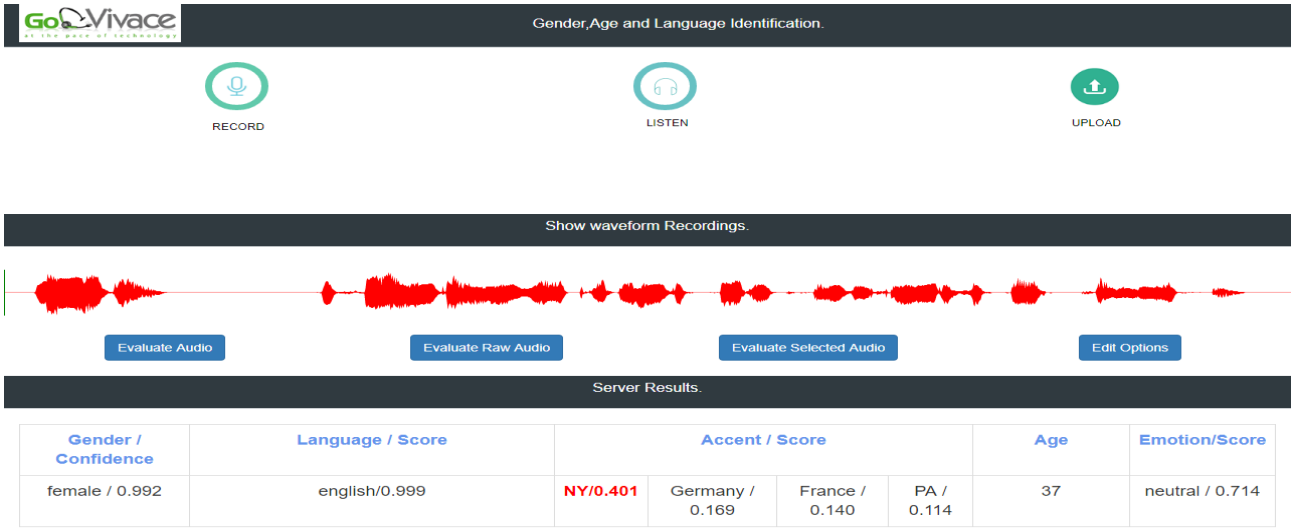

Figure 1: *Demonstration web page utilizing the websocket API*

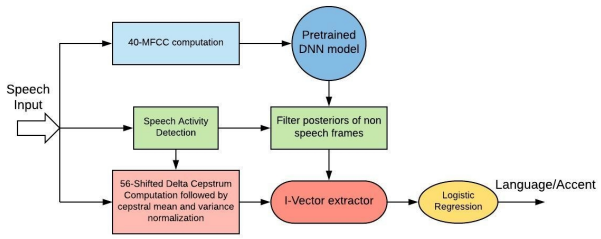

Figure 2: *System diagram for the gender, language and accent recognition subsystems*

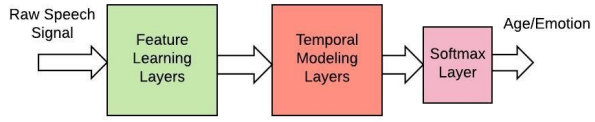

Figure 3: *Architecture of the age or emotion recognition subsystems*

The age and emotion identification engines are DNN system, which uses raw speech waveforms on the input layer of the DNN and generate categorical outputs. The DNNs have convolution layers based front-end [3] for learning features relevant to age/emotion. The DNN also have Long Short Term Memory (LSTM) cell based recurrent projection [4] for temporal modeling in the later layers. The architecture is depicted in Figure 3 and the emotion identification system uses the algorithm described in [5].

Table 3: *Percentage of recall per class for emotion recognition engine*

Neutral	Dominant
92.0	89.7

4. Performance of recognition engines and Conclusion

The performance of the recognition engines are individually evaluated on some fixed datasets with different utterance durations. We evaluate gender, language, accent and emotion recognition in terms of overall classification accuracy, which is shown in row 1 through 4 of Table 1. Language and accent recognition engine's performance becomes more meaningful when divided in terms of duration of utterances as shown in Table 2. Recall for individual emotion classes are shown in Table 3. The age recognition system's performance is computed in terms of mean absolute error (MAE) between actual and detected age, as shown in row 5 of Table 1, which is tested on 16.18 hours of data. Although the solutions are far from making no mistakes, they are much better than chance, and the current machines that are totally oblivious to speaker characteristics can still benefit from such a solution.

5. References

[1] M. Sarma, K. K. Sarma and N. Goel, "Language Recognition using Time Delay Deep Neural Network," $https://arxiv.org/pdf/1804.05000.pdf$, 2017.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transaction of Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[3] P. Ghahremani, V. Manohar, D. Povey and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *Interspeech 2016 – 17th Annual Conference of the International Speech Communication Association, September 8-12, San Francisco, CA, USA, Proceedings*, 2016.

[4] H. Sak, A. Senior and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in *Interspeech 2014 – 15thAnnual Conference of the International Speech Communication Association, September 14-18, Singapore, Proceedings*, 2014.

[5] M. Sarma, P. Ghahremani, D. Povey, N. Goel, K. K. Sarma and N. Dehak, "Emotion Identification from raw speech signals using DNNs," in *Interspeech 2018 – 19thAnnual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018.