Predicting Voice Elicited Emotions

Ying Li
Jobaline Inc.
620 Kirkland Way, Suite 208
Kirkland, Washington 98033, USA
ying.li@jobaline.com

Jose D. Contreras
Jobaline Inc.
620 Kirkland Way, Suite 208
Kirkland, Washington 98033, USA
jose@jobaline.com

Luis J. Salazar
Jobaline Inc.
620 Kirkland Way, Suite 208
Kirkland, Washington 98033, USA
luis@jobaline.com

ABSTRACT

We present the research, and product development and deployment, of Voice Analyzer™ by Jobaline Inc. This is a patent pending technology that analyzes voice data and predicts human emotions elicited by the paralinguistic elements of a voice. Human voice characteristics, such as tone, complement the verbal communication. In several contexts of communication, "how" things are said is just as important as "what" is being said. This paper provides an overview of our deployed system, the raw data, the data processing steps, and the prediction algorithms we experimented with. A case study is included where, given a voice clip, our model predicts the degree in which a listener will find the voice "engaging". Our prediction results were verified through independent market research with 75% in agreement on how an average listener would feel. One application of Jobaline Voice Analyzer technology is for assisting companies to hire workers in the service industry where customers' emotional response to workers' voice may affect the service outcome. Jobaline Voice Analyzer is deployed in production as a product offer to our clients to help them identify workers who will better engage with their customers. We will also share some discoveries and lessons learned.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - *data mining*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces -*voice I/O*; I.5.2 [**Pattern Recognition**]: Design Methodology - *pattern analysis*.

General Terms

Algorithms, Design, Experimentation, Human Factors.

Keywords

Voice Analysis, Predictive Modeling, Speech Signal Processing, Deployed System

KDD '15, August 11-14, 2015, Sydney, NSW, Australia.
ACM 978-1-4503-3664-2/15/08.
http://dx.doi.org/10.1145/2783258.2788619

1. INTRODUCTION

This paper presents Jobaline Voice Analyzer, a deployed system that analyzes human voice data and predicts listener emotions elicited by the paralinguistic elements of the voice [9].

The motivation for developing this technology at Jobaline is to automate the screening process of hiring workers for service industries. Jobaline is the leading mobile and bilingual hourly job matching marketplace and training network. Hourly job workers represent approximately two-thirds of the U.S. labor force and more than 50 million people are hired every year in this job category. The hiring of these hourly job workers is typically done through a manual process conducted by human resource personnel to match workers with jobs. Many hourly jobs can not be filled in time even though employers spend nearly one billion hours pre-screening job applicants. To overcome this extreme low efficiency, Jobaline automates the pre-screening and matching of applicants to jobs. The automation includes automated phone interviews whereby applicants record their answers to a set of interview prompts.

At this writing, Jobaline has processed over 700,000 job applications in either English or Spanish with millions of voice clips recorded by job candidates of different cultures, education levels, genders and ages. We resort to the power of data science and predictive modeling to analyze these voice clips to further facilitate the automation of recruiting processes. If we can identify applicants whose voices can elicit positive emotions in a listener, thanks to certain characteristics possessed by the voice, in real-time during the automated interview, our system can rank them higher in the job matching process. This in turn optimizes for best matching between workers and jobs.

We have thus researched and developed a voice data analysis system and built data mining models that can predict human emotions elicited by paralinguistic elements of the voice interview records. Our system utilizes only the paralinguistic elements (how things are said) of a voice and does not involve the lexical contents (what is being said). This has multiple benefits, including privacy preserving and fair hiring across cultural backgrounds.

Our main contributions are: 1) the framing of the problem of predicting listener emotion response to voice by paralinguistic elements; 2) a set of system modules that process and extract features representing voice cues; and 3) a set of data mining models and analyses that derive insights and make predictions about listener emotions in response to voice clips. The prediction results were verified by independent market research, and the system modules and prediction models are deployed into production. In particular, our prediction models quantify, for any given voice clip, the likelihood that a listener may associate certain emotions such as "engaging". The utilization of only

paralinguistic (i.e., non-lexical) elements enables us to predict listener response to "how things are said" regardless of what is being said.

One exemplary application of this technology is in the services industry, where hiring workers who can connect with their customers and keep their customers engaged is critically important. Some examples of these jobs include telemarketers, retail store clerks, frontline employees at a quick serve restaurant, or front desk associates at a hotel. By analyzing job applicants' voices with Jobaline Voice Analyzer, companies can predict which job candidates will likely perform better in sales and marketing or customer-facing positions at restaurants, hotels, or call centers. This provides an objective input into recruiters' hiring decision process.

This paper is organized as follows: Section 2 provides some basic definitions and background knowledge about voice data, and a high level overview of the state of the art of speech affect analysis. Section 3 presents our voice analyzer system: from the raw data, to feature extraction, to modeling, and deployment. Section 4 presents a case study where we built models which, given a voice clip, predict the likelihood of an emotion of "being engaged" generated in the listener. We also present validation of our prediction results through third-party market research. In Section 5, we share some discoveries from analyzing job applicants' voices and lessons learned. And lastly, in Section 6, we conclude with business benefits and outline next steps in our research and development.

2. BACKGROUND

For our goal of predicting emotion responses to voices, we envision a framework for research and development to encompass the below elements:

- an "emotion taxonomy" that includes all the emotion responses we want to predict such as "feeling engaged", "feeling soothed", or "trustworthiness"
- a general understanding of voice data that could provide insights for mapping voice characteristics to a feature space of finite quantified dimensions that can be computationally constructed
- a machine learning and data science experimentation framework that will enable us to build the best performing models

Among the many emotion categories and definitions researched and employed by the research fields, we found the framework articulated by "FEELTRACE" [2], screenshot in Figure 1, is most compelling for our purpose in the sense that it is more inclusive of all emotions we want to predict and it describes various emotions by finite quantifiable dimensions (see Figure 1). Whereas other research papers cover only a small set of "strong emotions", such as "fear", "happy", or "angry". The finite dimensions such as negative vs. positive, active vs. passive, could assist us in mapping emotion characteristics to and measure them by known voice paralinguistic features.

Combining emotion research with data science and engineering is a relatively new research and development area [10]. Our focus of predicting listener emotion response to voices would need to employ domain knowledge related to the processing and analysis of human voice data, as well as machine learning and data mining methodologies and techniques. The remainder of this section presents a brief introduction of affects associated with the human voice (Section 2.1), a high level survey of the state of the art in classifying emotions associated with voices (Section 2.2), and an overview of paralinguistic features and their use in existing works (Section 2.3).

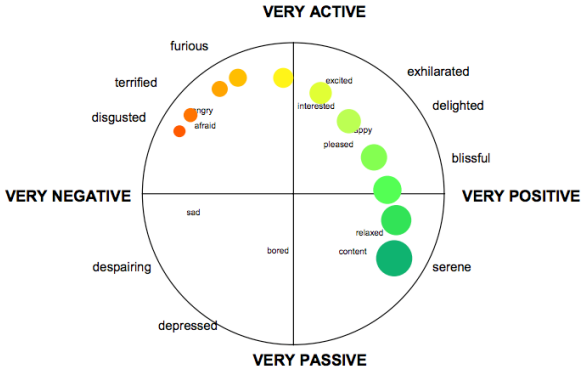


Figure 1. Representation of emotional states [2].

2.1 Affects Related to Human Voice

Attention to the affects of human voice probably existed as long as human existed. Systematic study of affects of human voice dated back centuries ago; Kreiman et al [6] provides a comprehensive survey on the studies of perception of voice quality. Research and studies on the affects in and from human voice can be found in subject areas such as theories of emotions for Information Retrieval [8], Affective Computing (e.g., MIT Media Lab Affective Computing Group), and Speech Communication (e.g., International Speech Communication Association) [14, 16, 17]. Table 1 lists some attributes associated with human voice that have been the subjects of research.

Table 1. Example Attributes Associated with Human Voice

Categories of Attributes	Definition	Examples
speaker trait	characteristics that are permanently associated with speaker	age, gender, personality, likeability
speaker state	attributes of speaker that change over time	affection, deception, interest, intoxication, stress
speaker acoustic behavior	non-linguistic vocal outbursts during speech	sighs, yawns, laughs, cries, hesitations
acoustic affect	non-linguistic attributes carried in the voice	sounds pleasant, cheerful, trustworthy, deceitful
elicited listener emotion	listener reactions immediate upon hearing a speech clip	feels energized, happy, joyful, annoyed, agitated

2.2 Classification of Affects Associated with Voice

There is a rich repertoire of scientific research in speech signal processing and analysis that make classifications of affects based

on human voice data. The majority of these research works can be regarded as addressing two sets of goals. One is to recognize the presence of and/or to classify the type of personality traits intrinsically possessed by the speaker. These personality traits can be independent of, or in relation to, when the speech was made. The other set of goals focuses on recognizing the presence of and classifying the types of emotions carried within a speech clip or the context out of which the speech clips arise.

Table 2 lists some previous research works that aimed at classifying speaker traits or presence of emotions in speech.

Table 2. Classification of Speaker Traits or Speaker Emotions.

Speaker Type	Speaker Prompt	Classification Objective
Parkinson's disease patient	portray emotion	presence of emotion in speech speaker trait [18]
Actor	portray emotion	presence of emotion in speech [12]
Actor	portray personality	speaker personality (e.g., Big Five) [11]
Consumer phone recording	customer seeking service and support from commercial service providers	presence of emotion in speech clips acoustic correlates observable with emotions [5]
Medical call center dialog	real life situation without artificial prompt	recognition of speaker emotion (anger, fear, relief, sadness) [3]

Jobaline Voice Analyzer is particularly focused on predicting the elicited emotions based on paralinguistic features of voice clips. We have not come across research works aimed at predicting/classifying listener emotions elicited by voices.

2.3 Paralinguistic Features of Voices

The modeling of the anatomy and physiology of human voice production have been studied in the domain of speech signal processing with established paradigm and methodologies [13]. We list some definitions related to analyzing paralinguistic features in Table 3.

Converting audio signals into data features typically involves:

- Sampling in time domain and frequency domain
- Extracting FFT and energy in frequency domain and information from the time domain signal
- Extracting feature vectors in the time and frequency domain

Generally, frequency and energy variation over time are the major cues used to analyze emotions in audio samples. A number of methods, from mathematical transformations to slicing audio samples into shorter snippets, turn direct acoustic signals into pitch contours, from which statistics can be extracted and analyzed for correlations with emotions.

Table 3. Definitions of some voice features.

Concept	Definition	Data Representation
amplitude	measurement of the variations over time of the acoustic signal	quantified values of a sound wave's oscillation
energy	acoustic signal energy representation in decibels	20*log10(abs(FFT))
formants	the resonance frequencies of the vocal tract	maxima detected using Linear Prediction on audio windows with high tonal content
perceived pitch	perceived fundamental frequency and harmonics	formants
fundamental frequency	the reciprocal of time duration of one glottal cycle - a strict definition of "pitch"	first formant

Existing research on detecting speaker emotions from voice clips has demonstrated that speaker emotions, and the intensity of those emotions, can be recognized from acoustic data with varying accuracy depending on the types of emotions. Pitch has been used as one major cue for emotion recognition. Intensity and speech rate have also been used. Typical transformations include turning a voice clip into pitch contours and measuring various statistics such as max, min, standard deviation, and time-window averages, on the whole clip or on snippets of the clip.

Some features that have proven effective for recognizing speaker emotions include:

- Fundamental frequency and its statistics such as min, max, mean and standard deviation over time
- Pitch contour
- Speech signal amplitude
- Frequency spectrum energy distribution
- Durations: proportion of pauses, duration of syllables, syllable rate, and total duration

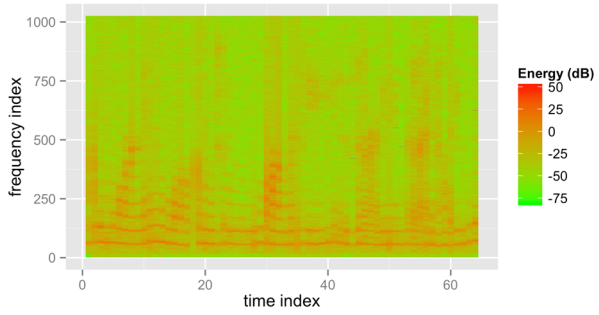
In addition, some research also demonstrated associations between voice features and emotions expressed in voices. For example, researchers found that the presence of anger in voice clips was associated with a rise in fundamental frequency and amplitude, whereas despondency was associated with a decreased syllable rate [5]. Researchers have experimented with novel acoustic features and have found that they outrank "classic" features for affect recognition tasks [4]. The use of paralinguistic features has also been demonstrated effective in assisting other features to further disambiguate affects.

The list of features presented at INTERSPEECH competitions, which focused on recognizing speaker emotion and likability through paralinguistic features [15, 16] provided a starting point for our feature space construction for our voice analyzer.

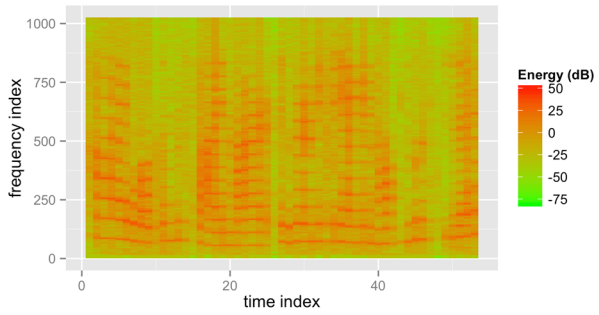
3. JOBALINE VOICE ANALYZER

Intuitively, our common day experience tells us that predicting emotions that might be elicited when we hear a voice clip is a feasible task. For instance, when we hear a voice, we can tell whether the voice makes us feel "being engaged", "at ease", or "soothed". Figure 2 shows the spectrograms of two sample voice clips from job applicants responding to the interview prompt,

"*Greet me as if I am a customer*". One can clearly notice the energy level difference between the two voice clips. The spectrogram on the top (Figure 2(a)) is from a voice clip that would make listeners feel much less engaging than the clips depicted on the bottom (Figure 2(b)).



(a) A non-energetic voice.



(b) An energetic voice.

Figure 2. Spectrogram of sample voice clips.

At Jobaline, we focused our study on the types and instances of paralinguistic features and their effectiveness in classifying voice recordings. We will describe our system in steps and provide descriptions about the components of our system.

3.1 Overview

Our task is to analyze and model speech affect based on paralinguistic correlates in acoustic data. The desired outcome is a system capable of predicting listeners' emotion response to voice clips elicited by the paralinguistic aspects of the voice clips. The building blocks of our system are drawn in Figure 3. The resulting Jobaline Voice Analyzer accomplishes the tasks through the following steps:

1. Record and sample raw voice clips
2. Extract audio features that represent voice cues
3. Construct data feature space suitable for applying data mining and machine learning algorithms
4. Build models using various algorithms for unsupervised learning, and supervised and semi-supervised learning
5. Engineer scalable data processing pipelines that process voice clips and generate prediction scores by applying the prediction models

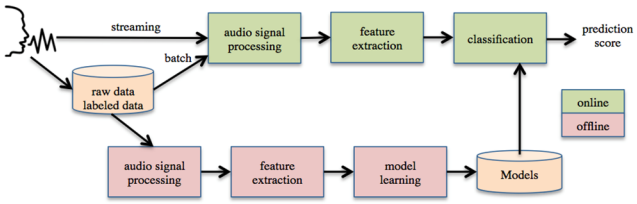


Figure 3. Jobaline Voice Analyzer building blocks.

3.2 Raw Data and Metadata

Our raw data are voice clips from job applicants recorded as audio signals in wave format. Jobaline has served more than one-half million job applicants, where each job applicant may record voice clips as responses to various interview prompts. Typical interview prompts are:

- *Greet me as if I am your customer.*
- *How would you describe excellent customer service?*
- *Tell me about a time you handled an angry customer.*
- *Tell me about a time when you were able to diffuse an escalated customer situation without transferring to a supervisor or other department.*
- *Briefly explain your experience on this type of job.*

The metadata associated with voice clips include job categories for which the applicants were applying and the interview prompts for which the applicants were asked to respond to by the employers. The metadata were used to filter voice clips for building models according to groups of interview prompts. The content of metadata is not used for modeling because we wanted our models to be purely based on voice data.

3.3 Preprocessing of Audio Signals

Our preprocessing tasks involves:

1. Converting audio signal into data in time domain and frequency domain
2. Filtering out voice clips that are unfit for modeling and analysis

Our audio signal processing component transforms voice clips in wave format to the following data elements:

- Short-term Fast Fourier Transform per frame
- Energy measures in frequency domain per frame
- Linear Prediction Coefficient in frequency domain per frame

The result of preprocessing is stored in a json file for each wave file. Our feature space for modeling is subsequently built based on these data elements (more in Section 3.4).

Voice clips that are not suited for modeling or scoring are those that:

- Are too short for any data elements to be meaningful
- Contain too much background noise, leading/trailing noise, or other types of noises

Since our voice clips are free-form speech recorded from job applicants, they do not have a uniform length when they arrive at our system. Figure 4 shows the distribution of voice clip length. We discard voice clips shorter than two seconds because they do not provide enough evidence of qualifications for employers to screen for further information.

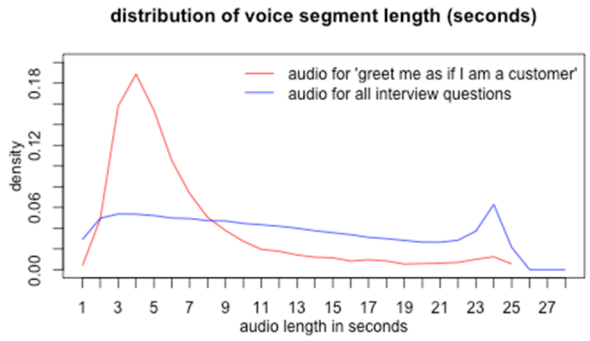


Figure 4. Distribution of voice clip lengths in seconds.

3.4 Feature Space Construction

We experimented with feature construction based on the following dimensions and combinations:

- Signal measurements such as energy and amplitude
- Statistics such as min, max, mean, and standard deviation on signal measurements
- Measurement window in time domain: different time size and entire time window
- Measurement window in frequency domain: all frequencies, optimal audible frequencies, and selected frequency ranges
- Distance metrics: dynamic time warping and Euclidean
- Algorithms: hierarchical clustering, k-means and complete

Combining our explorative analysis with learnings from existing researchers in speech signal processing and speaker emotion classification, we focused on voice energy features and constructed feature space using various statistical measures of energy attributes. We analyzed feature selection against clustering algorithms to determine the effectiveness of the features and to guide us in the final selection of features.

3.5 Model Building

Our modeling goal is to be able to construct prediction models that, given a voice clip, predict listeners' emotion response to it.

In modeling listener emotional response to voice clips, we cannot expect the availability of the absolute ground truth as we have yet to have a mathematical formulation of mappings from voice to emotion. We can jump start with the conventional approach of collecting training data through human labeling (labeled data), however we also note that the labelers' sense of emotion may vary and thus could potentially introduce hidden bias in the labeled data.

When given a set of voice clips without prior knowledge of quantified ground truth or definitive verified examples that represent ground truth, we first utilize unsupervised learning techniques to help us find patterns in unlabeled data. We then constructed training data sets based on evaluation of clustering results combined with independent manual labeling of individual voice clips. The labeled data are then fed to supervised learning routines to build the prediction models.

3.5.1 Unsupervised Learning

We ran clustering analysis on the paralinguistic features of the voice clips (as listed in Section 3.4). We experimented with:

- Various clustering algorithms such as hierarchical clustering and k-means
- Various distance metrics employed within the clustering algorithms, such as dynamic time warping and Euclidean

The clustering results were evaluated based on:

- Cluster quality measurements such as compactness, good separation, connectedness, stabilities, etc.
- Manual validation, including visual inspection of the clusters and/or aural inspection of the voice clips in each cluster, to determine whether the clustering is meaningful

We ran cluster quality measurements on multiple clustering algorithms and multiple choices for the number of clusters to determine which clustering algorithms yield the best results and the number of clusters that are most appropriate for the data. Figure 5 gives an example of those results, which indicates that hierarchical clustering with five clusters gives the best clusters in this particular example.

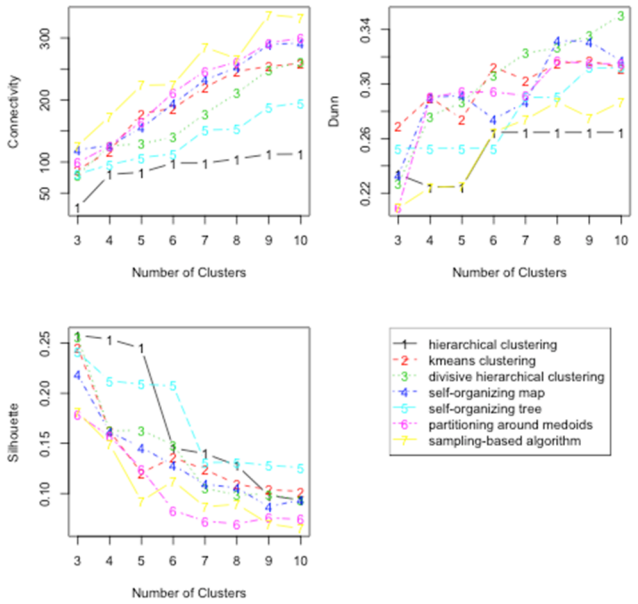


Figure 5. Cluster quality measurement by number of clusters.

We visualized the clustering results by displaying the centroids of clusters by average energy in frequency domain across voice clips in each cluster. Figure 6 shows the results of one such clustering run.

We listened to the voice clips that were clustered together to evaluate whether the voice clips from the same cluster did indeed generate similar listener effect. Based on our listening test, we detected reasonable similarities within clusters and dissimilarities between clusters, and they also mapped to a scaled positive vs. negative response.

For the example shown in Figure 6, we noted that two distinct clusters emerged: one for highly energetic voice clips and one for very low energy voice clips. This indicates that the voice clips and the corresponding extracted feature data have reasonable differentiating power to map the voice clips into clusters that might correlate with listener responses. We combine clustering results with human labeling to produce labeled data for training prediction models.

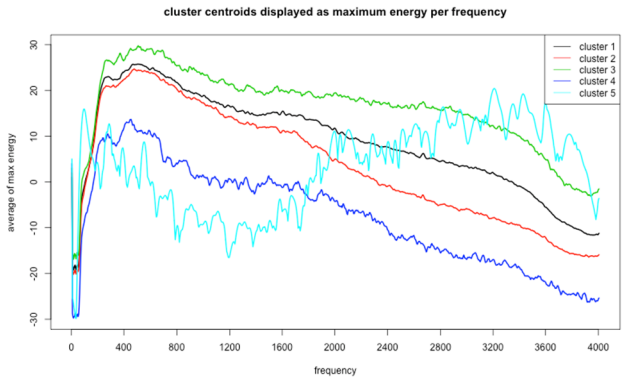


Figure 6. Clustering based on energy features.

3.5.2 Supervised and Semi-Supervised Learning

In building our prediction models, we experimented with various supervised learning algorithms, including:

- Logistic regression
- Support Vector Machine
- Random Forest

where the training data is a combination of clustering results and human labeling. We also experimented with models to predict binary outcome (positive vs. negative) and numerical scores for further classifying listener emotions.

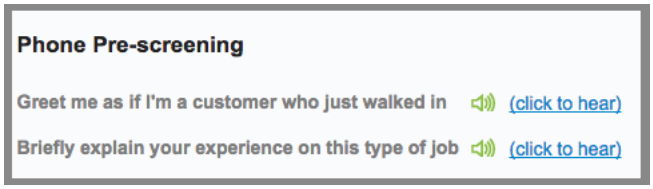
Our experiments show that the predictive models built on top of clustering insights and iterative feedback from listeners produces meaningful results. We will provide detailed accuracy analysis in Section 4 when we present our prediction model for emotion response of "feeling engaged".

In addition to prediction models using the convention methods like SVM and Random Forest, we also conducted experiments with an unsupervised and semi-supervised learning algorithm, called Kodama, which performs feature extraction from noisy and high-dimensional data [1]. The output of Kodama includes a dissimilarity matrix from which we can perform clustering and classification. In Figure 7, we project voice clips onto two-dimensional space from the multidimensional scaling of the Kodama dissimilarity matrix. The voice clips' emotion response label is overlaid with color ("turquoise" for positive response, and "red" for negative response). We can see a fairly good separation of the two types of responses in training data. We use experiments like this to validate the direction of research investigations, and to lead to the final selection of training data and corresponding features to be used in model training.

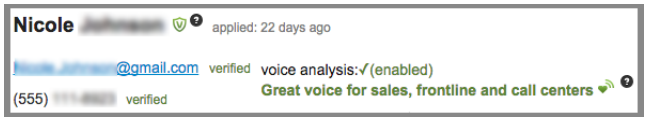
3.6 Deployed System

The Jobaline Voice Analyzer system is a production system deployed in a cloud infrastructure. The functional components as described in Figure 3 are all deployed to production: voice data arrive at our system after a job applicant records their answers to interview questions; a prediction score is produced by the system; and the result is stored in a database for downstream consumption, such as ranking of applicants. The main architecture consists of: Hadoop clusters; audio signal preprocessing components developed in MATLAB and then converted to Java running on the clusters; various model training and prediction routines written in R.

Figure 7. Overlay emotion label on Kodama dissimilarity.



(a) Interview questions and links to voice clips.



(b) Voice analysis conclusion at lower right.

Figure 8. UI screen shots of how voice analysis is surfaced.

Figure 8 provides two screen shots of Voice Analyzer results that are surfaced to the recruiter UI as in the deployed system. Figure 8(a) shows the UI where a recruiter as a Jobaline user can access the voice recordings left by the job applicants as answers to their interview questions. Once a voice clip went through our system, a notice and flag will show on the UI, as shown in lower right of Figure 8(b), to indicate whether the voice would score high as to elicit a positive response from customers for the types of jobs the candidate is applying for. The blur in Figure 8(b) is added for the purpose of this paper to mask out the candidate's personal data.

At the time of this submission (February 2015), Jobaline had more than 700,000 job applicants on file, processed more than two million voice clips, and Jobaline Voice Analyzer was already available as product offering to clients.

Our work has been focused on the paralinguistic aspects of the voice data. We do not analyze the lexical contents of the voice clips. The result of Voice Analyzer is one additional data point to the recruiting process, not a hiring decision. The use of lexical content of the candidate's answers to the interview questions in the recruiting process is left to the human recruiter.

4. PREDICTING LISTENER EMOTION OF FEELING ENGAGED

We present a case study for predicting listener emotion responses to voices for the purpose of assisting employers in screening job applicants. The business goal is to enable better matching of workers to jobs that require interaction with customers and keeping them engaged. Examples of such jobs are telemarketer, retail store clerk, frontline employee at a quick serve restaurant, or a front desk associate at a hotel.

We incorporate the thinking from the emotion presentation framework by Cowei et al [2] (Figure 1). Instead of taking any specific categorization and definition of one particular emotion (such as "happy") and building a model for it, we chose to start by predicting a positive response vs. negative response. A positive response could be one or multiple perceptions of a "pleasant voice", "makes me feel good", "cares about me", "makes me feel comfortable", or "makes me feel engaged".

While being novel in predicting listener emotion responses, we take the learnings from classifying speaker traits and speaker states by previous research works as guidance. In particular, the effectiveness of certain speech paralinguistic features in identifying the presence of and further classifying speaker emotions helps us decide what features to use to construct our feature spaces.

4.1 Prediction Model for Engaging Effect

We used the methodologies and system components described in Section 3 to obtain the prediction model and deployed the model to production. The production system was built and upgraded in two versions by the time of this writing. We experimented with SVM for version 1 and Random Forest for version 2.

The collection of voice clips that answer the interview prompt of "*Greet me as if I am a customer*" are used for building the models for engaging effect. The training data was obtained by the combination of unsupervised learning and human manual labeling and inspection.

We have achieved accuracy in the range of 76% to 90% as measured by cross-validation for binary classification (positive vs. negative on "feeling engaged", "energetic", etc.). Here we report one model built with random forest algorithm, it has an AUC value of 0.918, the ROC for binary classification performance as shown in Figure 9, and some other performance measures as listed below:

```
##               Accuracy : 0.8605
##                 95% CI : (0.7689, 0.9258)
##    P-Value [Acc > NIR] : 5.762e-07
##             Sensitivity : 0.8182
##             Specificity : 0.8868
##          Pos Pred Value : 0.8182
##          Neg Pred Value : 0.8868
```

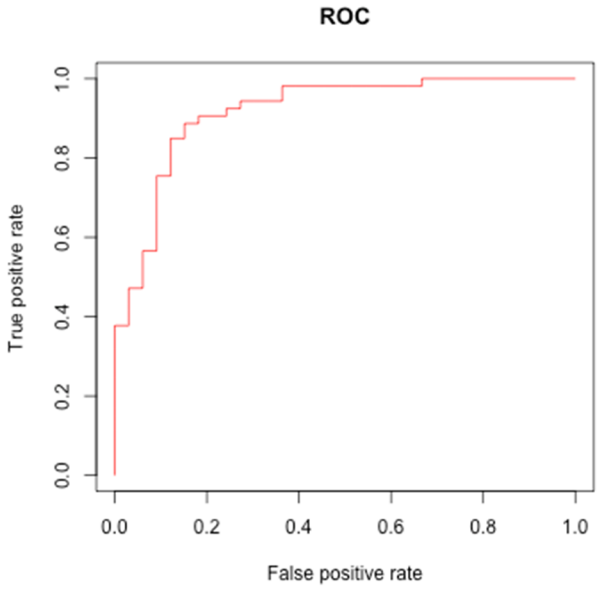


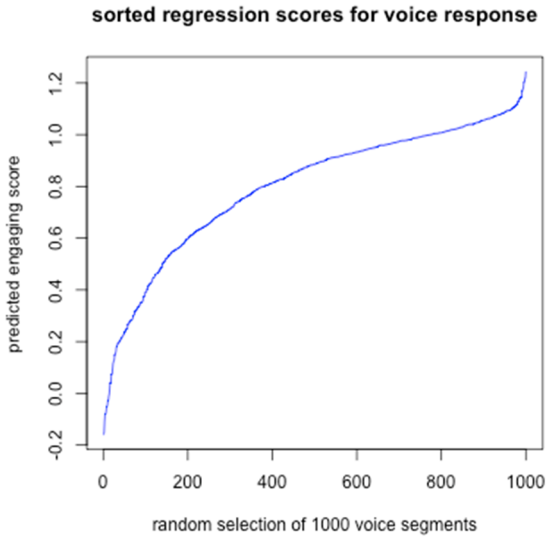
Figure 9. Cross validation ROC for deployed model.

4.2 Prediction Scores for Downstream Applications

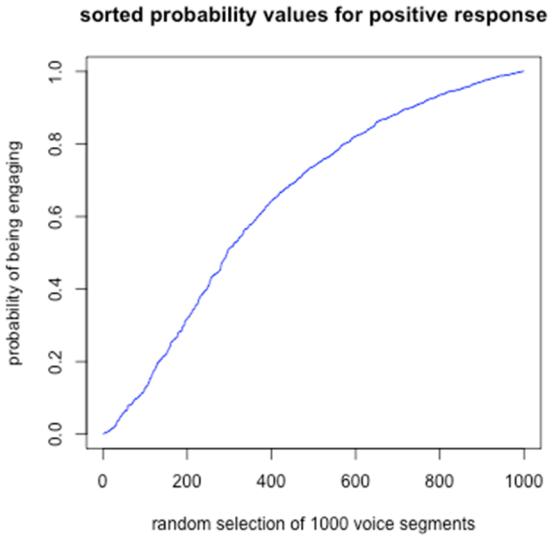
The downstream applications of the prediction scores need to be based on a set of well thought out principles. As we iteratively build out our prediction engines for voice elicited emotions, our system is comprised of models built with different training datasets, different predictive modeling algorithms, and different tuning parameters. The prediction scores will have different meaning and implications. Figure 10 shows the spread of the predicted scores on 1000 randomly selected, unseen voice clips, by a support vector regression model (Figure 10(a)), and a random forest classification model (Figure 10(b)). The two sets of scores are directionally parallel, but have to be treated differently, as the model in Figure 10(a) predicts the engaging score and the model in Figure 10(b) produces the probability of positive response to a voice clip.

Each voice clip is very rich in what it expresses; listener response to voices is complex depending on the listener's context such as cultural and demographical factors. Thus, even within the same model, say the random forest model (Figure 10(b)), when applying the scores in downstream applications such as ranking, a more reasonable way to use the prediction scores is to bucketize voice clips directionally, accordingly. As we will continue to improve on all aspects of our efforts, such as feature extraction, feature transformation, combination of training datasets, modeling algorithms, and their parameters, the absolute scored values will undergo changes until a global scaled normalization is found satisfactory.

We can also observe important differences in model behavior from Figure 10. The SVM model in Figure 10(a) tends to be more biased towards predicting positive response than the random forest model. One reason for this behavior is that the training dataset was imbalanced due to our limited labeled training data at the earlier stage of our engine development. We moved our production system to the version 2 model that is random forest based with more balanced training dataset.



(a) Scores from v1 model of support vector regression.



(b) Scores from v2 model of random forest classification.

Figure 10. Sorted prediction scores by two models for 1000 randomly selected voice clips.

4.3 Market Research for Evaluation and Model Refinement

To get third party opinion on our production models, we used market research with a representative U.S. demographic sample (by geographic region, income, ethnicity, gender and age) to manually validate the results.

Jobaline commissioned two independent research firms to validate the output of our prediction models. This validation was done by surveying a representative sample of U.S. residents balanced by geographic region, income, ethnicity, gender and age, comparing their subjective evaluation of voice clips with our prediction results. The survey takers matched the U.S. census in geography, age, ethnicity, income, and gender, at a 95% confidence interval and with a 5% margin of error. Survey responses have to be triple verified in order to be included in calculating agreement between the survey taker and our prediction results. Figure 11 describes the verification process in detail.

The market research concludes that, when asked how the sound of the recorded voice clips made them feel, 75% agreed with the Jobaline Voice Analyzer prediction of an interesting or engaging voice.

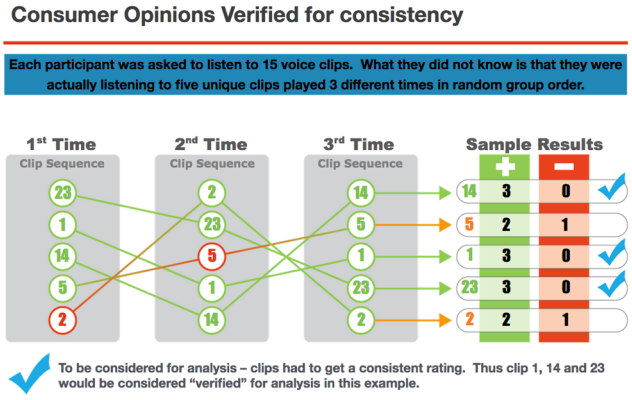


Figure 11. Only those survey responses that were triple-verified are used in validation.

Our market research data could also provide insights to feedback on our feature construction and modeling work. Some key takeaways are:

1. The demographic characteristics of the listener matter. For example, younger listeners (18-29 years old) or people in lower income brackets of less than $29K/year have more strict criteria of how they sense pleasant or engaging. The practical implication is that algorithms can be fine-tuned to the age range of the target audience the workers will speak to. Thus, a retailer that caters to a younger demographic might need individuals with voice characteristics different than a retailer that caters to an older demographic.

2. There is a significant drop in emotional response to voices of similar characteristics when the listener is exposed to voice clips longer than five seconds. More research has to be done to correlate this to various demographics, but this could affect things like defining the optimal length of customer greeting for a telemarketing or customer service firm, or a retailer based on the demographic they serve.

3. No positive or negative correlation was found between the emotion elicited and the age, ethnicity (accent), or education level of the speaker.

4. A slight bias of the respondents toward female voices was noticed. Voices with similar characteristics, but from a female speaker ranked on average 11% better than voices from male speakers. This provides additional input for fine-tuning the algorithms; not based on gender, but on the unique attributes of the voice.

5. DISCOVERIES AND LESSONS LEARNED

Since the launch of Jobaline Voice Analyzer, we have received inquiries from media and interests in collaboration from academia. Here we want to share some of the discoveries we've made while working with our voice data and lessons learned.

5.1 Differences in Voice Characteristics

It is an established knowledge that men and women have different fundamental frequencies in their voices. On average, the fundamental frequency for men is about 115 Hz and 220 Hz for women [8]. We selected one male voice and one female voice from our voice data set and show, in Figure 12, their average fundamental frequencies. The fundamental frequencies of the male voice (top graph) show differences between two interview questions (red vs. blue), whereas the fundamental frequencies of the female voice (bottom graph) are well mixed within the same range for different interview questions.

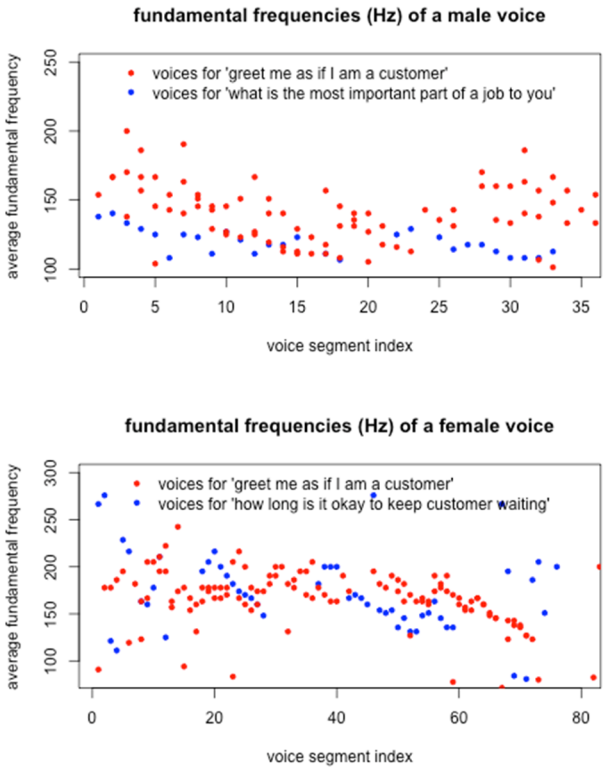


Figure 12. Fundamental frequencies for one male voice (top graph) and one female voice (bottom graph).

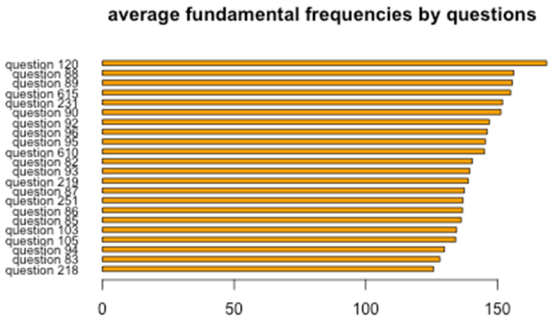


Figure 13. Differences in fundamental frequencies for different interview questions.

Taking all job applicants on all interview questions, we have observed differences in voice characteristics when job applicants answer different interview questions. Shown in Figure 13, the fundamental frequencies from voice clips on the question, "greet me as if I am a customer" (question 120 in Figure 13) are much higher than the fundamental frequencies on question, "tell me about the last time you performed a similar job" (question 218).

5.2 Layering Classification Models by Voice Characteristics Helps Improve Accuracy

Taking into consideration the differences in voice characteristics, we experimented with layering a decision tree by features related to fundamental frequency on top of the random forest classification model. We are able to improve the classification accuracy by nearly 10%. The layered model will need to go through further market research validation as described in Section 4.3 before we put it into production.

5.3 Anomalous Voice Characteristics

A voice paralinguistic feature can sometimes be critical for one classification task and be noise for another. For instance, "silence intervals" can be a useful feature for predicting soothing response when the voice is intentional for keeping a steady rhythm, but it can also be noise for predicting engaging effect when they are indications of low energy in the voice.

Dealing with voice signal often involves trade-offs between time domain and frequency domain. We have learned lessons to be mindful about using statistical features taken over one domain. For instance, when we use the maximum energy in the frequency domain over time as the main feature, it generally works well for voice clips that are of similar time length, but it has potential to overrate a voice clip that is very long in time. Therefore we need to take into account such limitations in the models/features when we use the scores from the model in a production environment where the voice clips were generated in free form. When we removed "silence intervals" in our voice feature construction, it helped remove noise, but it also generated a boundary case where a mostly silent clip actually gets high engaging score.

6. BUSINESS BENEFITS AND FUTURE RESEARCH

With Jobaline Voice Analyzer, employers will be able to identify service workers whose voices will better engage with their customers. Workers and employers will benefit from a shorter hiring cycle, thanks to the automated nature of the process. The deployed Jobaline Voice Analyzer system is starting to provide more quality input into our data science research and will enable more robust product features from Jobaline in the future.

We have a number of exciting future projects that will explore the many rich features of the Jobaline voice database and tap into vast repositories of data science methodologies and techniques. We have started work on building models for predicting emotion response of feeling "calmed" when exposed to voices that the listener feel is soothing. Our current findings indicate that we need to incorporate more paralinguistic features, such as cadence and emotion changes over time.

Over the longer term as we deploy data science on more types of emotion responses, as depicted in Figure 1, we hope to synthesize and establish a set of research methodologies on paralinguistic voice features for modeling listener response for broader application areas such as human-machine interaction, opening the doors for creating artificial voices for machines and allowing for better communication by enabling an emotional machine-listener connection.

7. REFERENCES

[1] Cacciatore, S., Luchinat, C., Tenori, L. 2014. Knowledge discovery by accuracy maximization. In *Proc. Natl. Acad. Sci., USA, vol. 111 no. 14, 5117-5122.*

[2] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M. 2000. FEELTRACE: an instrument for recording perceived emotion in real time. In *ISCA workshop on speech and emotion*, Northern Ireland, pp 19–24.

[3] Devillers, L., and Vidrascu, L. 2006. Real-life emotions detection with lexical and paralinguistic cues on human call center dialogs. *INTERSPEECH*.

[4] Fernandez, R. 2004. A Computational Model for the Automatic Recognition of Affect in Speech. PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA.

[5] Forsell, M. 2007. Acoustic Correlates of Perceived Emotions in Speech. Master Thesis in Speech Communication, Royal Institute of Technology. KTH.

[6] Kreiman, J., Van Lancker-Sidtis, D., and Gerratt, B.R., 2005. Perception of Voice Quality, in The Handbook of Speech Perception, Ed. Pisoni, D.B. and Remez, R.E., Blackwell Publishing, 338-362.

[7] Kreiman, J., Sidtis, D., 2011. Foundations of Voice Studies. Wiley-Blackwell.

[8] Lopatovska, I. and Arapakis, I. 2011. Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interactions. *Information Processing & Management*. 47(4), 575-592.

[9] Mullor, M., Salazar, L., Li, Y., and Contreras, J. (Jobaline, Inc., USA) 2015. Matching and Lead Prequalification Based on Voice Analysis. US Patent Application #14532600.

[10] Picard, R.W. 2010. Emotion research by the people, for the people. *Emotion Review,* Volume 2, Issue 3 (July 2010)

[11] Polzehl, T., Moller, S., and Metze, F. 2010. Automatically assessing personality from speech. *2010 IEEE Fourth International Conference on Semantic Computing (ICSC).*

[12] Polzin, T. S., and Waibel, A. 1998. Detecting emotions in speech. *Proceedings of the CMC*.

[13] Quatier, T. F. 2002. Discrete-Time Speech Signal Processing: Principles and Practice.

[14] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., et al. 2010. The INTERSPEECH 2010 paralinguistic challenge. *INTERSPEECH*.

[15] Schuller, B. 2011. Voice and speech analysis in search of states and traits. *Computer Analysis of Human Behavior,* 227-253.

[16] Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., et. al. 2012. The INTERSPEECH 2012 Speaker Trait Challenge. *INTERSPEECH*.

[17] Weiss, B. and Burkhardt, F. 2012. Is 'not bad' good enough? Aspects of unknown voices' likability. *INTERSPEECH*.

[18] Zhao, S., Rudzicz, F., Carvalho, L. G., Márquez-Chin, C., and Livingstone, S. 2014. Automatic detection of expressed emotion in Parkinson's disease. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*