CLASSIFICATION OF VOICES THAT ELICIT SOOTHING EFFECT BY APPLYING A VOICED VS. UNVOICED FEATURE ENGINEERING STRATEGY

Ying Li, Kevin Mueller, Jose D. Contreras, Luis J. Salazar
Jobaline Inc., Kirkland, Washington 98033, USA
{ying.li, kevin.m, jose, luis}@jobaline.com

ABSTRACT

This paper introduces a novel approach of classifying voices that elicit a soothing effect on listeners from a domain knowledge inspired application of feature engineering. In particular, we utilize the characteristics of voiced vs, unvoiced speech in order to build a more accurate feature set. Large sets of training data are prepared and disciplined feature selections are conducted. Our final classifier achieved 86.84% classification accuracy of cross validation and evaluations by unknown listener population via crowdsourcing have rates of agreement with the classification model range from 80% to 90%. The technologies are deployed into Jobaline products to help service companies identify hourly-job workers whose voice can elicit soothing effect on customers.

Index Terms— Voice Analysis, Paralinguistics, Emotion Recognition, Voiced vs. Unvoiced, Acoustic Feature Engineering

1. INTRODUCTION

In this paper, we classify voice clips that generate a soothing effect on a listener vs. those that do not. Classification models are used to predict the likelihood that a voice segment would elicit a calming feeling in the listener. The prediction engine is deployed into Jobaline's hourly job matching marketplace and training network, where automated phone interviews record applicants' answers to a set of interview questions. The prediction models are applied to the recorded voice clips and the prediction scores are surfaced to recruiters as an input into the job matching process.

Existing work on classification of affects associated with voices can be viewed, at a very high level, as addressing two sets of research objectives [9]: 1) to detect the presence of and/or classify the types of personality traits intrinsically possessed by the speaker [25]; 2) to recognize the presence of and/or the types of emotions or acoustic events carried within a voice clip or the context out of which the voice clip arises. The speaker personality traits can be independent of, or in relation to, when the speech was made, therefore, can be further distinguished as speaker trait (e.g., age, gender, personality) classification [22, 23, 25, 26, 28] and speaker state (e.g., affection, intoxication, stress) classification [8, 19, 22]. The analysis of emotions or events carried within a voice can be about the acoustic behavior (e.g., sighs, hesitation, laughs) and acoustic affect (e.g., pleasant, deceitful, cheerful) classification [32]. The work presented in this paper is a continuation of our earlier work reported in [16], where we presented Jobaline Voice Analyzer [20] and a machine learning approach to predict voice elicited emotions in a listener, particularly focused on predicting the likelihood of a listener feeling of being engaged upon hearing a voice clip.

In our prediction models, we intentionally do not use meta data we have on job applicants, nor the lexical content of the voice recordings, and only utilized paralinguistics. Paralinguistics have been widely used in many tasks related to emotion recognition and in speech data processing in general [27], they generally consist of many features that depend on voiced speech such as fundamental frequency and formants, which played important roles but the measurement of which could be incorrect during unvoiced speech [2, 3, 15, 30].

Our work differs from previous research mainly in three aspects: 1) we focus on predicting how a listener feels upon hearing a voice, instead of the emotion of the speaker. 2) The emotion of feeling soothed or calming has not been as widely studied as the "strong" emotions (e.g., angry, fear, happy, sad) in previous research; 3) we apply a voiced vs. unvoiced feature engineering strategy to get a more accurate feature set; 4) we utilize paralinguistics in a "big data" fashion such that crowd sourced labels and machine learning and predictive modeling are scalable to voice collections, of sizes that are rarely reported in previous research initiatives.

Our contributions are: 1) a classification model for classifying voice clips as having a soothing effect on listeners with cross validation accuracy of 86.84%; 2) a set of paralinguistic features computed on a very large set of voice clips from a very large set of speakers that is representative of the 76 million hourly-job workers in the United States; 3) an application and verification of current feature extraction methodologies such as voice vs. unvoiced segmentation; 4) validation by using feedback loop for building prediction models on arbitrarily large sets of voices, by utilizing Amazon Mechanical Turk to cross-validate classification models and iterate the training of the models to improve accuracy.

The remainder of this paper is organized as follows: Section 2 motivates our research by business needs and

introduces the voice data collected at Jobaline used for modeling. Section 3 presents paralinguistic features we experimented with. Section 4 describes our feature selection and model selection and Section 5 presents results.

2. PROBLEM STATEMENT AND DATA SET

A voice that elicits soothing effect in listeners is often desired at service entities such as retail customer service lines and call centers to help ensure customer satisfaction. For example, a call center may wish to identify workers whose voices elicit a soothing effect on customers and place them in suitable positions.

Our research problem is: given a voice clip, predict the likelihood of listener feeling soothed by the voice upon hearing the voice, based on the paralinguistics of the voice clips.

Jobaline's automated interview process asks job applicants to answer a set of interview questions determined by employers. Sample interview questions 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*.

At the time of this submission, Jobaline has processed over 1 million job applications in either English or Spanish with millions of voice clips recorded by job applicants. They are voices of different cultures, education levels, geographical locations, genders and ages as diverse as the general US hourly-jobs work force. The voice recordings are of the following specification: 8000 sampling rate, mono channel, and 16 bits per sample wave format. The duration of the voice clips is unconstrained as job applicants freely decide how much they record for each interview question.

Our exploratory analysis on the Jobaline interview voice clips revealed that the clips for different interview questions exhibit systematic differences in some paralinguistics, such as mean fundamental frequencies, which may be attributed to the nature of the interview prompt, such as interactive vs. monologue. We decided to use the voice clips for a single interview prompt for each model, and in the rest of this paper, we report our work on voice clips for answering the interview question of "*how would you describe excellent customer service*".

3. PARALINGUISTIC FEATURES

Adopting the terminology of organizing paralinguistic features from the comprehensive survey [27], our acoustic features and functionals explored for this research are listed in Table 1. The calculations of these paralinguistic features are mainly implemented in Python. We also used MATLAB, PRAAT [4], and R package Seewave [29] to conduct additional experimentation and analysis. The differences in the paralinguistic measurements are significant enough to impact the classifier accuracies. We devised the voiced vs. unvoiced feature engineering strategy to obtain more accurate measurements of paralinguistics.

Table 1. Acoustic and functional features experimented.

Features	Functionals
F0 (Voiced)	mean, std,skew 1st derivative, max/min
F1,F2,F3 (Voiced)	mean,std, kurtosis first derivative
MFCC	std, skewness
Auto-Correlation	skewness
HNR	mean,std,skewness
ZCR	skewness,mean
LPC Coefficients	std,mean,skewnss
STFT	mean,std

3.1. Voiced vs. unvoiced strategy

Our motivation for the voiced vs. unvoiced (V/UV) strategy was that incorrect calculations of voiced features could cause error when applying the large number of feature functionals required. For example, fundamental frequency has been shown to correlate with intonation [10], which from manual inspection was found to be an important acoustic to consider in characterizing a soothing voice.

Voiced speech can be considered a segment with approximately constant frequency tones of some duration, whereas unvoiced speech is composed of non-periodic and random sounds caused by air passing through a narrow constriction of the vocal tract. Common examples of voiced speech and unvoiced speech are vowels and consonants respectively [15]. Although a difference in interpretation exists, for our purposes we also consider silent segments to be unvoiced speech.

Our work can be seen as a data science variant to common pitch trackers, such as [30], which applied a two-pass normalized cross-correlation (NCCF) to a down sampled signal in order to build (V/UV) hypothesis sets. One advantage of our implementation is less computational complexity, at the expense of attaining labeled data. Our work differs from other data science solutions such as [3], as it combines more sophisticated feature extraction strategies such as [2,17], allowing for less over fitting in small datasets. *Algorithm 1* outlines the procedure used to obtain and process the voiced and unvoiced feature sets

Algorithm 1. Voiced vs. unvoiced feature engineering

1. Compute the average signal energy across all frames
2. Compute maximum signal energy from frames
3. For each frame
   a. Calculate all frame-level paralinguistic features
   b. Apply training (V/UV) classifier
   c. If voiced, update voiced features
4. If no frame is voiced, impute missing values for voices features

3.2. Classifier for voiced vs. unvoiced segments

The data consisted of a training set hand labeled from 10 randomly selected frames for 40 clips in our database. Because accurate computations of the fundamental frequency were our primary motivation, we considered a frame to be voiced if the fundamental frequency existed and could be accurately computed. We used a small feature set

of what we considered to be the best representation of a (V/UV) signal, including: HNR, ZCR, signal energy and the ratio of high (>2000Hz) to low (<=2000Hz) frequency components. We applied a random forest classifier and our final model achieved a cross-validation accuracy of 93%.

3.3. Acoustic Features

We used 50ms windows with 50% overlap and applied pre-emphasis and a Hamming window for preprocessing. For voiced features, we calculated the first three formants and the fundamental frequency. Other low level descriptors included: ZCR, signal energy, HNR, LPC coefficients, and the autocorrelation function. The first three formants were calculated by finding the peaks of a 32-order LPC spectral envelope. We also calculated the first 42 MFCC coefficients and STFT independently of our (V/UV) implementation and windowing procedure. Finally we computed the unvoiced ratio from the output of our (V/UV) classifier [27, 33].

4. CLASSIFICATION OF SOOTHING EFFECT

We followed the typical steps of supervised learning to build classification models:

- Prepare training data and collect class labels
- Conduct feature selection and model training
- Inspect model results with cross validation and validation in the wild
- Iterate to improve model accuracy

4.1. Preparation of training data

We prepared training data as a binary class label on each voice clip, 1 for the clip that makes listeners feel soothed when they hear the voice , 0 for listener not feeling soothed. We used two ways to obtain label data: 1) using our own R&D team members to label voice clips; 2) using crowdsourcing on Amazon Mechanical Turk (AMT) to label voice clips. Data labeled by our own team has the advantage of high quality on each clip and consistency across labelers, it has the limitation of narrow demographics of the audience and hard to scale to large collection of clips. Labeling by AMT workers can give us wide audience and scale to arbitrary large collection of clips that is only constrained by monetary budget.

Data labeling by uncontrolled and untrained workers poses potential bias of subjectivity and inconsistency, especially so when the subject of labeling is emotion where there are multiple schools of thoughts and even more definitions and representations [18, 24,13, 14], pursuant to different viewpoints on dimensions, such as perception of voice quality, emotion taxonomy, emotion process, and so on. Our attempt to limit labeling inconsistency are two mechanisms: 1) a descriptive instruction to the labelers; and 2) inter-rater agreement. Table 2 shows instructions that AMT labelers see when they label each clip. We address the inter-labeler consistency by asking at least 10 labelers on each clip and use only those clips that received the same label from more than 80% of labelers. For this study, we did not filter on intra labeler consistency, which we may introduce in the future depending on the demand and cost. We have mechanism for identifying fraudulent labelers.

Table 2. Instructions to labelers

Listen to the following clip, does he/she make you feel something from column **A** or something from column **B**	
A	**B**
• I find it soothing	• I do not find it soothing
• I find it calming	• I do not find it calming
• I find it relaxing	• I do not find it relaxing
• Makes me feel at ease	• It does not makes me feel at ease
• I feel as if he/she cares about me	• I feel as if he/she does not care about me

4.2. Feature and model selection

Because our feature space is constructed in a brute-force manner, it is important to conduct feature selection in order to reduce the size of the feature set and achieve higher model accuracy. We adopt a wrapper approach to feature selection [11]: the machine learning algorithm we chose to use is Random Forest [5] as implemented in R [6], and we correlation threshold to search the feature space for feature selection. *Algorithm 2* outlines the feature selection procedure. Figure 1 depicts the feature selection search space (step 3 of *Algorithm 2*) by model accuracy over the threshold used for filtering highly correlated features, with (V/UV) strategy (top) and without (bottom) respectively.

Algorithm 2. Feature selection with learning using randomForest

1. Compute the correlation matrix of the feature space
2. Define a threshold set to be a set of values between 0 and 1
3. For each value in threshold set
   a. filter out the variables that have correlation value above threshold
   b. train randomForest model on remaining features and calculate cross validation accuracy measures
4. Select the threshold that gives maximum accuracy by the desired accuracy measurement (e.g., overall accuracy, balanced accuracy, false positive, etc.) and output the set of features after filtering out features above this threshold

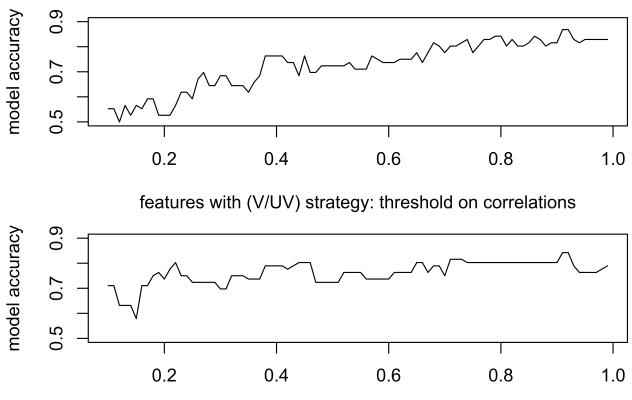

Figure 1. Model accuracy corresponding to threshold for filtering out correlated features.

Table 3. Top 20 features by importance measure of mean decrease in accuracy.

With V/UV strategy	Without V/UV
skewmfccs0	meanMFCC9
stdmfccs2	skewnessMFCC1
meanstFFTdB109	mean162
meanstFFTdB128	meanMFCC11
meanmfccs0	sdMFCC3
meanstFFTdB166	skewnessMFCC2
meanstFFTdB36	sd243
meanmfccs4	sd3
stdmfccs0	meanMFCC5
skewmfccs1	meanF0
skewZCR	meanMFCC8
meanmfccs9	skewnessMFCC4
stdstFFTdB181	mean8
stdstFFTdB242	sdMFCC2
stdmfccs6	mean3
percentFalseVoicedvsUnvoicedFrames	sdMFCC10
stdmfccs4	sdMFCC8
meanf0	skewnessMFCC5
meanstFFTdB1	sdF0
meanmfccs1	sd182

4.3. Cross-validation and evaluation in the wild

Our final model was a randomForest model built with the feature set selected by *Algorithm 2* (Section 4.2), with a training set of 775 labeled voice clips, 309 as positive and 466 as negative. Top 20 features are listed in Table 3 by order of variable importance measurement of mean decrease in accuracy. Cross validation was carried out by splitting labeled dataset into training and testing by various proportions such as 70/30, or 90/10. The best overall accuracy by cross validation is 86.84%. Various accuracy measurements are listed in Table 4.

We conducted multiple validations-in-the-wild over the course of our research. The validation set is either 500 randomly selected unseen voice clips, or top 25% and bottom 25% from 1000 randomly selected clips that are scored by our final soothing classification model and ranked according to class probabilities. Validation sets were sent to AMT to be evaluated according to the same instruction as the training data were labeled. Comparing the AMT results with model scored results, the rates of agreement range from 80% to 90%. Accuracy measurements are listed in Table 4.

4.4. Feedback loop for improving model accuracy

The newly labeled clips not only serves the purpose of validation at wild, but also serve as additional labeled data to train the next iteration of the model. The motivation for sending top 25% and bottom 25% for validation in the wild is to reduce waste by increased probability of inter-rater agreement when we feed the new labels back into the learning process to train the next iteration of models for improved accuracies. This cycle was repeated regularly and stable or slightly improved model accuracy has been obtained over time.

Table 4. Cross validation accuracy and validation of agreement between AMT labelers and prediction model.

accuracy measure	CV without voiced vs. unvoiced	CV with voiced vs. unvoiced	validation in the wild
Accuracy	0.8421	0.8684	0.9
95% CI	(0.7404, 0.9157)	(0.7713, 0.9351)	(0.8466, 0.9396)
P-Value [Acc > NIR]	0.000006767	4.572E-07	<2e-16
Kappa	0.6492	0.7181	0.7936
Sensitivity	0.9783	0.9348	0.9327
Specificity	0.6333	0.7667	0.8553
Pos Pred Value	0.8036	0.86	0.8981
Neg Pred Value	0.95	0.8846	0.9028

5. CONCLUSIONS AND FUTURE WORK

Our work has produced classification models for soothing effect on listeners with accuracies ranging from 86% to 90%. We have established validity and feasibility to construct learning models based on arbitrarily sized training sets of voice clips limited only by crowdsourcing cost. As of this submission, we have 35,276 labels on 3276 clips, out of over 1 million unseen voice clips. We plan to keep iterating on our model with steady growth of training labels.

This work of predicting voice effects on listeners is only a small step towards appreciating the full complexity in human's judgment of vocal quality, such as the ability to sooth listeners. We observed a preference of female voices over male voices exhibited in our AMT labeled data, which is consistent with findings in [1]. We experimented with building layered models with a crude gender classifier, with no significant results of improvement. We may utilize other advanced approaches such as [7, 31] to continue to improve our model accuracy.

Our validation in the wild tested on the agreement of our model with how listeners feel out of listeners that were from unknown populations, different from [7] that dealt with testing on data from population different from training. We observed that validation in the wild sometimes achieves higher accuracy than cross validation. We plan to investigate whether this is caused by choice of thresholds on class probabilities alone or whether over-fitting is at play.

The success in utilizing the (V/UV) strategy for model accuracy improvement encourages several different avenues for future work. For example combining the (V/UV) strategy with similarity measures for vowel detection between voiced frames [23] or syllable nuclei [22] could allow for better measures of speech rate. In addition, further models of intonation could be built by utilizing metrics of pitch peaks allowing for a more condensed feature set.

We have utilized feature scaling with some success, however accuracy lift varied. We have experimented with PCA with no noticeable improvement. We plan to improve our feature selection approach for potential additional lift in model robustness and accuracy.

6. REFERENCES

[1] Babel, M., McGuire, G., King J., "Towards a More Nuanced View of Vocal Attractiveness", PLoS ONE 9(2): e88616. doi:10.1371/journal.pone.0088616, 2014.

[2] Bagshaw, P.C., Hiller, S.M., and Jack, M.A., "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching", Proceedings of Eurospeech, 1993.

[3] Bendiksen, A. and Steiglitz, K., "Neural Networks for Voiced/Unvoiced Speech Classification", *Proceedings of ICASSP*, pp.521 -524, 1990.

[4] Boersma, P., and Weenink, D., "PRAAT: doing phonetics by computer", Version 5.4.08, http://www.praat.org , 2015.

[5] Breiman, L., "Random Forests", *Machine Learning*, volume 45, issue 1, pp. 5 - 32, 2001.

[6] CRAN, "Breiman and Cutler's random forests for classication and regression", https://cran.r-project.org/web/packages/randomForest/index.html .

[7] Deng, J., Xia, R., Zhang, Z., Liu, Y., and Schuller, B., "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition"., *INTERSPEECH*, 2014.

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

[9] El Ayadi, M.M.H., Kamel, M.S., and Karry, F., "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognition*, volume 44, pp 572 – 587, 2011.

[10] Gussenhoven, C., The Phonology of Tone and Intonation, Cambridge University Press, 2004.

[11] Guyon, I. and Elisseeff, A., "And introduction to variable and feature selection", *Journal of Machine Learning Research*, volume 3, pp. 1157 - 1182, 2003.

[12] Hirschberg, J. et. al., "Distinguishing deceptive from non-deceptive speech", *INTERSPEECH*, 2005.

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

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

[15] Lee, J.K., Yoo, C.D., "Wavelet speech enhancement based on voiced/unvoiced decision", *The 32nd International Congress and Exposition on Noise Control Engineering*, Seogwipo, Korea, August 25-28, 2003.

[16] Li, Y., Contreras, J. D., and Salazar, L. J., "Predicting voice elicited emotions", *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, 2015.

[17] Liu, J., Zheng, F., Deng, J., and Wu, W., "Real-time pitch tracking based on combined SMDSF", *Eurospeech*, 2004.

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

[19] Mitra V., Shriberg E., "Effects of feature type, learning algorithm and speaking style for depression detection from speech", *ICASSP*, 2015.

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

[21] Pavol Partila, P., Voznak, M., Mikulec, M., and Zdralek, J., "Fundamental frequency extraction methods using central clipping and its importance for the classification of emotion state", *Information and Communication Technologies and Services,* 10(4), 2012.

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

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

[24] Scherer, K. R., "What are emotions? And how can they be measured?" in *Social Science Information*, Volume 44, Issue 4, pp 695-729, 2005.

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

[26] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., Narauanan S., "The INTERSPEECH 2010 paralinguistic challenge", *INTERSPEECH*, 2010.

[27] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., Narauanan S., "Paralinguistics in speech and language – State-of-the-art and the challenge", *Computer Speech & Language*, Volume 27, Issue 1, pp 4 – 39, January 2013.

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

[29] Sueur, J., Aubin, T., and Simonis, C., "Seewave: a free modular tool for sound analysis and synthesis", *Bioacoustics*, 18, pp 213 – 226, 2008.

[30] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", In *Speech Coding and Synthesis*, edited by Klein, W. B., and Palival, K. K., 1995.

[31] Vogtm T., André, E., "Improving automatic emotion recognition from speech via gender differentiation", *Proc. Language Resources and Evaluation Conference*, Genoa, 2006.

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

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