DISCRIMINATION OF SUNG VOWELS AND BREATHY OR PRESSED PHONATION USING MFCC FEATURES

Nicholas Bone

Department of Computer Science, Western Washington University, USA

ABSTRACT

Feedback based on expert aural discernment is vital for beginning singers. Most automatic speech recognition (ASR) systems are concerned with discerning intent rather than giving feedback to the subject. Instead of intent-based labels, this research uses perception-based labels from three independent human judges. We extract mel-frequency cepstral coefficients (MFCCs) from recorded song and train simple multilayer perceptrons (MLPs) for each of three aural discernment tasks: vowel classification, breathiness rating, and pressedness rating. Inter-judge agreement analysis with cross-validation demonstrates expert-level performance with good generalization to novel singers. In the breathy and pressed tasks, mean squared error (MSE) is lower for the machine-human comparisons than for the human-human comparisons. Vowel classification accuracy is around 70% for the humans and 60-80% for the machines, depending on how truth is defined.

***Index Terms*—** MFCC, breathy voice, pressed voice, vowel classification, inter-judge agreement

1. INTRODUCTION

Singing analysis has been an understudied area in audio signal processing. It is well known that beginning singers lack sufficient aural discernment to tell right from wrong in their vocal production [1], which is one reason voice teachers are so important. The goal of this research is to develop a "pocket voice teacher" that can serve as an expert ear to give feedback on certain important qualities of singing. Unlike speech recognition, which is typically concerned with discerning the intended meaning of an utterance, this research is concerned with the perception of the sound to a listener. There are myriad vocal qualities that voice teachers listen for and give feedback on to their students. Here we pick two: diction (specifically vowel discrimination) and voice quality (flow phonation versus breathy or pressed).

A vast majority of ASR research is concerned only with recognition, and not evaluation. Training corpora such as TIMIT have single ground-truth labeling with no conception of ambiguity, and phones are typically folded into a small set of classes that cannot represent the diversity of pronunciation in the audio input [2]. Computer-aided pronunciation training (CAPT) for second language learning is the main exception, as its purpose is to provide evaluative feedback. However, like most ASR systems, CAPT evaluation models are typically trained from single-truth corpora (both [3] and [4] used TIMIT for learning English phones).

Most research into voice quality discrimination uses speech data with fixed classes (e.g., breathy, modal, and pressed) and intent-based labeling. [5] compared a variety of features for distinguishing breathy vowels in Gujarati speech; they emphasized loudness measure, but also demonstrated good discrimination with spectral features. [6] used MFCCs to train a MLP to estimate the glottal open quotient (OQ), which is not directly measurable from the sound signal (it's usually inferred from an electroglottograph) but is such a good predictor of voice quality that much research has gone into finding robust correlates.

One of the few published attempts to train a machine to discern breathy or pressed phonation in singing, [7] used iterative adaptive inverse filtering (IAIF) to estimate glottal waveform features, for classification with intent-based labels from a single singer. They dismissed spectral features like MFCCs as unsuitable, since phonation modes are governed primarily by the glottis and not the vocal tract.

This research uses perception-based labeling from three independent human experts. Learning from perception-based labels allows for discrimination of subtle differences in the sound that would be trained away in intent-based labeling. Using labels from multiple human experts allows analysis of ambiguity among the input sounds and measuring inter-judge agreement for more comprehensive performance analysis.

In a further departure from previous voice quality research, we use regression rather than classification to allow fine-grained evaluation of breathy and pressed phonation.

2. DATA COLLECTION AND PROCESSING

This section describes our procedures for collecting the sound recordings, segmenting the recordings into sample units, extracting feature vectors, and labeling the samples.

2.1. Recording Procedure

All our input data come from original recordings of ten volunteer singers, collected using a handheld digital recorder set on a table a few feet away from the singer in an otherwise quiet office with a "normal" amount of reverb (no apparent "liveness" due to bare walls and such). Each recording features one singer with no other instruments or background noise. We collected six recordings from each singer:

Vowels baseline – The singer sings the seven Italian vowels /i e ɛ ɑ ɔ o u/, on G (G3 for men and G4 for women) sustaining each for about a second.

Vowels scales – The singer sings the same sequence of vowels, but this time each one sung on a rising C major scale (*do, re, mi, fa, sol, la, ti, do*), pausing for breath in between vowels, but not while singing the scale.

Italian song – The singer sings an Italian art song of their choice, given the constraints that it should be legato and feature a range of vowels and pitches. The goal here is to collect some "more realistic" data for testing the system. Of course, there is still no accompaniment, and the recording is in a controlled environment, but it's a step toward realism.

Breathy scales – The singer repeats the "vowels scales" exercise, but sings as breathy as they can. The singer is instructed to maintain *mezzo-forte* dynamics to counteract the natural tendency to associate "breathy" with "*piano*" and just sing softer.

Normal scales – This is an exact repetition of the "vowels scales" exercise, to reacquaint the singer with their "normal" (hopefully "flow-like") sound, and provide a fair baseline of comparison for the "breathy" and "pressed" data.

Pressed scales – Repeat the same exercise once more, but this time make it sound as pressed as they can (without hurting themselves, and while still sounding human). Again, the singer is instructed to maintain the same tempo and dynamics as for the previous two recordings.

Our goal was to get clear, sustained vowels and representative examples of breathy, normal, and pressed phonation at a range of pitches. We encouraged the singers to take it slow, and allowed them to redo any recording they were not satisfied with.

The six recordings together yielded about 4-6 minutes of raw data for each singer.

2.2. Feature Extraction

We used Praat[1] and custom scripts to extract feature points from the raw sound files. Because we were testing only on sustained vowels we used a window length of 0.25 seconds for the extraction and a time step of 0.02 seconds. The primary reason to use a wide window was to get smoother formant values, which facilitated segmentation (discussed in the next section).

We extracted 18 features at each time step: time offset (milliseconds from the beginning of the recording), intensity, pitch (F0), the first three formants (F1-F3), and the first 12 MFCCs. Only the MFCCs were used for training the vowel classifier and breathy/pressed regressors; the first six features were used in the automatic segmentation program. The time offset was also used along with the participant and recording IDs to uniquely identify each feature vector.

2.3. Segmentation

Our experiments required uniform-length, single-vowel sound samples. It was important to strip away the surrounding context from each sound sample so that the human judges' ratings would be based on aural discernment rather than contextual knowledge not available to the machine. We chose half-a-second as the length for the sound samples; there were plenty of sufficiently sustained vowels among the input recordings, and half-a-second is long enough for human ears to reasonably evaluate.

We created a PITCHVOWELSEGMENTER program to automatically extract half-second samples of sustained vowels from the 60 input recordings. This program used a sliding window along the feature vectors for each recording and identified and extracted ranges of voiced vowels (based on intensity and having defined formants) with reasonably constant pitch, F1, and F2. We used a 12% pitch tolerance to account for vibrato (+/- one semitone), and 15% F1 and 20% F2 tolerances, arrived at through experimentation. In some cases two consecutive samples were extracted from the same sustentation, but samples never overlapped.

This processing yielded, on average, about 300 samples for each participant (over 25 minutes of sample data total).

2.4. Labeling

Most previous work has used intent-based labeling, but we are specifically concerned with perception-based labeling. There is often a disconnect between the intended sound and the perceived sound, especially for beginning singers, so being able to give automatic feedback to the singer about how they're coming across is valuable. To this end we hired three expert voice teachers to label the data. We ran-

[1] www.praat.org

domly selected 1232 samples to label by vowel and 447 to label with breathy and pressed ratings, uniformly distributed among the participants. (We chose not to label all the data due to the time and expense, and to save our judges from going mad.)

We created a VOICELABELER web app to facilitate the labeling process. It presents the sound samples in a random order for each judge to mitigate contextual bias in the labeling. There are eight label options for the vowels: the seven Italian vowels plus a button for "none / mixed / unclear", which the judges were instructed to use for samples that did not sound clearly like one of the seven target vowels. For the breathy and pressed ratings there are two independent sliders (one each for breathy and pressed) with a 0-20 point scale so as not to impose an arbitrarily coarse structure on the ratings.

3. EXPERIMENTS

We trained simple multilayer perceptrons for all three learning tasks, with 12-dimensional input vectors (the MFCCs) with each dimension normalized to [-1, 1] across the example set. The vowel classifier has seven outputs (one per target vowel) and eleven hidden units. The breathy and pressed regressors have a single linear output and eight hidden units. For all tests we used sigmoid activation, a learning rate of 0.2 with annealing, momentum of 0.5, and a maximum of 100 training cycles (preliminary experiments on a subset of the training data showed these to be reasonably effective parameters).

The labeling was performed at the sample level, where each half-second sound sample comprises 25 data points (due to the 0.2s time step), all of which received the same vowel label or breathy/pressed ratings. We treated these data points independently for the purposes of training in order to limit the dimensionality of the input and because the learning targets are not time-dependent. To get the sample-level predictions for each trained model we aggregated over the 25 component data points. For the vowel classification task the predicted label is defined as $p = \arg\max_v \sum_{i=1}^{25} confidence_{v,i}$, where $confidence_{v,i}$ is the normalized output of the neuron corresponding to vowel v for the i^{th} data point. For the breathy and pressed tasks we used the mean rating among the sample's component data points (the 25 outputs are truncated to the range [0, 20] before averaging).

The three human judges often disagreed about the ratings, so there was no clear "ground truth" to use for training. For each learning task we ran several experiments with varying definitions of "truth":

3-Judge Vowels – This experiment trained only on samples for which at least two of the three human judges agreed on the vowel label. Samples with unanimous agreement were given twice the training weight as samples where one judge dissented. 1107 of the 1232 labeled samples met these criteria, with 630 having unanimous agreement.

2-Judge Vowels – For each pair of judges we trained on the vowel labels for which they agreed: 821, 805, and 741 samples qualified for these three experiments.

1-Judge Vowels – We used each judge's labels as the lone truth for these three experiments. As above, "unclear" samples were excluded, resulting in training sets of 1222, 1147, and 1119 samples.

3-Judge Breathy and Pressed – These two experiments (breathy and pressed each used the same set-up, but with different ratings) used the mean ratings from all three human judges as the target for training. All 447 rated samples were included, but those where the judges agreed were given higher weight using the formula $w(s) = \frac{3}{3 + \text{range}(s)}$, where $\text{range}(s)$ is the absolute difference between the minimum and maximum rating for sample s among the three judges.

2-Judge Breathy and Pressed – These six experiments were the same as the *3-Judge* experiments except they used the mean ratings from only two judges at a time.

1-Judge Breathy and Pressed – These six experiments used just one judge's ratings at a time as the target for training. Since there is no measure of confusion when the ratings are from a single judge, all samples were given equal weight.

For all experiments we used cross-validation, training on data from nine of the ten singers and testing against the unheard singer. Thus, the results are indicative of how such a system might generalize to novel singers.

4. RESULTS

Average inter-judge agreement for the vowel classification task was 65% over the 1232 labeled samples, ranging from 60% to 70% depending on which pair of judges are compared. 3-judge agreement was much more likely given 2-judge agreement (77%) than 2-judge agreement given the lack of 3-judge agreement (43%), which corroborates the feedback from the judges that some samples were much more difficult to classify.

Factoring out "unclear" labels, aggregate inter-judge accuracy was 71%, as illustrated by the confusion matrix of Table 1. The distribution is unsurprising given the proximity of these vowels, and that all singers and judges are native English speakers: the /ɔ/ vowel, rare in English, is easily confused with the neighboring /ɑ/ and /o/, which are also seldom heard in pure form in English speech. Aggregate accuracy of the *3-Judge Vowels* model tested on unheard singers was 65%, shown in Table 2, and showed a similar distribution. (Note that the maximum possible aggregate accuracy here is 85% given the judge disagreement.) Other

trained models performed similarly; Table 3 lists the accuracy scores of the three judges and the various trained models relative to different definitions of truth (aggregate accuracy is not shown, but is roughly the average of the A, B, and C columns).

Tables 4 and 5 show the root mean square error (RMSE) of the three judges and trained models for the Breathy and Pressed experiments, respectively, relative to the different truths described in the experiments section. The high inter-judge RMSEs reflect the difficulty of the labeling task, with "pressedness" being more difficult to discern than "breathiness" from isolated half-second sound samples. (Given the distribution of ratings, the expected RMSE would be about 9 if there were zero correlation; the actual correlation between the judges' ratings was about 0.6 for Breathy and 0.5 for Pressed.)

We tested against unheard singers to assess generalization potential, given that MFCCs are also effective features for speaker differentiation and therefore different singers will manifest the same vowel differently. For comparison we also trained using all singers' data and reserved just the Italian song samples for testing: the vowel classification accuracies were around 5 percentage points higher, slightly surpassing the human judges; the RMSEs for the breathy and pressed tasks averaged about 0.15 lower than the already better-than-human values in Tables 4 and 5. This confirms that MFCC patterns vary significantly between individuals, and also underscores the inconsistencies in human labeling of ambiguous data.

5. DISCUSSION

Perhaps our most important result is that "accuracy" is a misleading metric when applied to supervised learning with subjective labels from a single source. Consider the m_B result in column C of Table 3: 66% accuracy is not bad for seven-way classification, but seems far from perfect when compared to the implied ideal of 100%, considering that vowel recognition seems trivial for humans. However, accepting the difficulty of the task when non-aural context is stripped away, 66% accuracy is actually very good, and exceeds the 65% accuracy of Judge B from whose labels m_B was trained. 100% accuracy relative to single-source labels would amount to severe overfitting. Near-100% accuracy is a reasonable aspiration only for the subset of "very clear" samples on which all three judges agreed.

Our results on the Breathy and Pressed tasks show "better than human" performance. This, too, is somewhat misleading. Given that the human ratings are the source of truth, superhuman performance is achievable only to the extent that the humans were individually inconsistent in their ratings. Comparisons of the ratings and self-reports from the judges both suggest some inconsistent "guessing"

	i	e	ɛ	ɑ	ɔ	o	u
i	1102	112	0	0	0	0	6
e	112	594	201	0	0	1	0
ɛ	0	201	514	5	0	4	2
ɑ	0	0	5	908	161	66	11
ɔ	0	0	0	161	262	188	17
o	0	1	4	66	188	512	186
u	6	0	2	11	17	186	842

Table 1: Inter-judge confusion matrix for vowel classification.

	i	e	ɛ	ɑ	ɔ	o	u
i	1138	230	12	0	0	10	8
e	88	524	160	4	0	6	4
ɛ	2	246	538	56	10	32	38
ɑ	0	0	28	950	262	140	38
ɔ	0	0	0	94	156	150	20
o	0	2	16	70	186	454	196
u	4	20	34	36	36	190	788

Table 2: Confusion matrix of the *3-Judge Vowels* model predictions (rows) relative to the aggregate "truth" of the three human judges (columns). Values doubled for comparison with Table 1.

	ABC	AB	AC	BC	A	B	C
A	1	1	1	.850	1	.659	.716
B	1	1	.767	1	.719	1	.646
C	1	.783	1	1	.734	.606	1
m_{ABC}	.827	.740	.776	.790	.664	.620	.674
m_{AB}	.841	.760	.789	.791	.686	.624	.672
m_{AC}	.841	.762	.784	.787	.684	.615	.677
m_{BC}	.825	.742	.773	.783	.663	.615	.670
m_A	.808	.737	.745	.775	.647	.627	.657
m_B	.819	.735	.755	.773	.652	.616	.656
m_C	.837	.769	.772	.800	.678	.647	.673

Table 3: Accuracies of the three judges (A, B, C) and seven trained models (subscripts indicate the judges from whom they learned) relative to labels agreed upon by (all) the judge(s) listed at the top.

	ABC	AB	AC	BC	A	B	C
A	2.99	3.32	2.62	4.48	0	6.64	5.24
B	4.55	3.32	6.82	3.96	6.64	0	7.92
C	3.89	5.84	2.62	3.96	5.24	7.92	0
m_{ABC}	3.41	4.33	3.33	4.02	3.98	6.61	4.48
m_{AB}	3.70	4.69	3.47	4.26	4.25	6.92	4.45
m_{AC}	3.64	4.69	3.35	4.21	4.20	6.95	4.30
m_{BC}	3.58	4.52	3.43	4.17	4.13	6.77	4.50
m_A	3.73	4.56	3.60	4.37	4.10	6.84	4.77
m_B	3.49	4.38	3.38	4.13	3.98	6.68	4.56
m_C	3.72	4.47	3.70	4.37	4.09	6.73	4.93

Table 4: RMSE of the judges and models for the Breathy task relative to the (mean) rating of the judge(s) listed at the top.

	ABC	AB	AC	BC	A	B	C
A	5.06	4.32	3.86	7.59	0	8.64	7.73
B	4.30	4.32	6.45	3.10	8.64	0	6.20
C	3.68	5.52	3.86	3.10	7.73	6.20	0
m_{ABC}	3.96	4.44	4.32	4.80	6.27	6.12	5.29
m_{AB}	4.15	4.72	4.71	4.65	6.84	5.92	5.23
m_{AC}	4.16	4.76	4.65	4.69	6.80	6.03	5.19
m_{BC}	4.40	4.93	4.77	5.05	6.75	6.36	5.46
m_A	4.13	4.64	4.27	5.08	6.15	6.52	5.33
m_B	3.98	4.46	4.29	4.87	6.20	6.21	5.30
m_C	4.29	4.70	4.62	5.12	6.42	6.34	5.60

Table 5: RMSE of the judges and models for the Pressed task relative to the (mean) rating of the judge(s) listed at the top.

on the less-clear samples. However, some of the inter-judge disagreement is likely due to their different internal rubrics. This could potentially be mitigated using adaptive alignment to translate each judge's ratings to the same operational scale.

Performance in all three tasks was markedly improved with training exposure to the target singer. Further research is needed to determine whether a diverse enough training set could close this gap. An "enrollment" procedure could also be effective to adapt a pre-trained model to novel singers.

6. CONCLUSION

This research was specifically concerned with characteristics that can be discerned from narrow windows of sound, independent of the surrounding context. Humans are not used to processing brief excerpts of voice out of context; like most ASR systems, we learn to perceive *intent*, and we subconsciously resolve ambiguity based on contextual cues, non-aural stimuli, and background knowledge. Even expert voice teachers have difficulty with aural discernment when context is stripped away: all the judges expressed the experience of second-guessing themselves at times during the labeling process. It is no wonder then, if each judge sometimes disagrees with herself, that there is so much inter-judge disagreement. Certainly, if the samples had been in the context of a song there would have been much greater agreement among the humans as they all leveraged contextual knowledge not available to the machine, and correspondingly worse generalization of the learned models given apparently self-contradictory sets of training labels.

We've shown that expert-level performance is easily achievable on vowel and breathy/pressed discrimination tasks on recordings of singing in a controlled environment, using only MFCC features. The models are quick to train and simple enough to be employed in a real-time feedback system. The procedures described in this paper would likely be effective for a wide range of aural discernment tasks. Some other vocal characteristics important in singing—such as vibrato, onsets, and general diction (including phone clusters)—are temporally defined, and would therefore require more complicated models. This research did not assess robustness in the face of background noise (especially instrumental accompaniment) or different recording environments, though these would be important features in a "pocket voice teacher" application.

7. REFERENCES

[1] A. S. Bone, "Time Use, Strategic Behaviors, Technical Content, and Cognitive and Motivational Profiles in Collegiate Vocal Music Practice," Ph.D. dissertation, University of Washington, Seattle, 2011.

[2] C. Lopes and F. Perdigão, "Phoneme Recognition on the TIMIT Database," in *Speech Technologies*, InTech, 2011, pp. 285-302.

[3] K. W. Yuen, W. K. Leung, P. F. Liu, K. H. Wong, X. J. Qian, W. K. Lo and H. Meng, "Enunciate: An Internet-accessible computer-aided pronunciation training system and related user evaluations," in *Proc. Oriental COCOSDA*, 2011.

[4] Y. B. Wang and L. S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. ICASSP*, 2012.

[5] S. A. Thati, B. Bollepalli, P. Bhaskararao and B. Yegnanarayana, "Analysis of breathy voice based on excitation characteristics of speech production," in *Proc. SPCOM*, 2012.

[6] J. Kane, S. Scherer, L. P. Morency and C. Gobl, "A comparative study of glottal open quotient estimation techniques," in *Proc. INTERSPEECH*, 2013.

[7] P. Proutskova, C. Rhodes, T. Crawford and G. Wiggins, "Breathy, Resonant, Pressed–Automatic Detection of Phonation Mode from Audio Recordings of Singing," *Journal of New Music Research*, vol. 42, no. 2, pp. 171-186, 2013.