CAN VOICE QUALITY IMPROVE MANDARIN TONE RECOGNITION?

Dinoj Surendran, Gina-Anne Levow

University of Chicago
Computer Science Department
1100 E. 58th Street
Chicago IL 60637

ABSTRACT

We investigate several measures of voice quality (VQ) to improve tone recognition in Mandarin Chinese. We find that band energy measures such as Spectral Balance (Sluijter and van Heuven, 1996) work better than measures based on glottal flow estimation and harmonic-formant differences. We also determine a set of bands and measures that improve tone classification accuracy on broadcast news speech to 64.1% from 60.4% when added to a traditional pitch-duration-intensity set of features. Most improvement is for the neutral tone, for which the F score increases from 0.345 to 0.619.

Index Terms— Speech recognition, Speech processing, Feature extraction.

1. INTRODUCTION

Traditionally, acoustic features used to automatically recognize Mandarin tones are based on pitch, duration, and overall intensity. We wish to know if other acoustic cues can offer additional information, particularly cues that measure the 'strength of a syllable' in some sense.

Mandarin tones are typically defined in terms of targets specifying pitch height and contour: high level, mid rising, low, and high falling. The neutral tone does not have a well-defined target, falling on unstressed syllables and being contextually determined. Neutral tone is thus very poorly characterized and badly recognized by the typical pitch, intensity and duration features.

It is reasonable to believe that strength-based cues can help recognize neutral and possibly low tones, since syllables with neutral tone cannot be lexically stressed, and because low tones are sometimes produced with creaky voice [1].

There has been much investigation in the last ten years of Voice Quality (VQ), or how far away a segment of speech is from its modal form (an 'average' half-open half-closed setting of the vocal folds). It has proved useful for various recognition tasks, such as detecting phrase boundaries in English [2] and Swedish, pitch accent in German [3] and prominence in English [4].

2. TASK DESCRIPTION

We wish to know if VQ cues can aid tone recognition when added to traditional acoustic cues. PID68 is a set of 68 features based on pitch, overall intensity, and duration [5]. Durational features included the length of the syllable and its rhyme, and the number of voiced frames in them. Pitch features included a 6-point contour and its difference, gradients of various parts of the contour, and the mean, maximum, standard deviation, etc, of pitch during the syllable. Pitch features adjusted by the mean pitch of the previous syllable were also used. Likewise with intensity, except that the adjustments used were by the mean intensity of both neighbor syllables.

In each of our experiments, we fixed a dataset of Mandarin broadcast news speech and a classification algorithm, and computed classification performance when using PID68, and when using PID68 plus a d-dimensional vector of VQ features, where d varied with the set of features considered.

Datasets were subsets of stories from the Voice of America Mandarin TDT 2 corpus [6] that had been automatically segmented, force aligned, and manually spot-checked [7].

For classification we used a 1-versus-1 ensemble [8] of Regularized Least Squares linear binary classifiers [9] with Platt-scaled outputs [10] that produces probability estimates as predictions[1]. For each syllable, the classifier estimated the probability that it had each of the 5 tones.

In each binary classification subproblem, we have N D-dimensional training examples $x_1, \ldots, x_N \in \mathbb{R}^D$ with ± 1 labels $y_1, \ldots, y_n \in \{-1, 1\}$ and find weights $b \in \mathbb{R}, w \in \mathbb{R}^D$ such that if $z_i = w^T x_i + b$, the sum $\sum_{n=1}^{N} (y_i - z_i)^2 + \lambda(w^T w + b^2)$ is minimized. We used $\lambda = 1$ for all cases.

3. MEASURES OF VOICE QUALITY CONSIDERED

As there is no standard measure for VQ, we tried several. Each feature for a syllable was Z-normalized by its distribution over all syllables in the same news story; speakers changed across stories, not within.

[1]The C++ scalable classification package we implemented for our experiments is available at http://people.cs.uchicago.edu/∼dinoj/nafla .

3.1. Glottal Flow Estimation

Some VQ measures are based on estimating glottal flow during speech using idealized templates (of varying shape) of glottal air flow. With a triangular template, **OQa** is the fraction of the period that is spent opening the glottis, and **ClQ** is the fraction of the period that is spent closing the glottis. Both are lower when the voice quality is higher [11]. With a rectangular template, the fraction of glottal closing time is called the Normalized Amplitude Quotient (**NAQ**) [12]. Other related measures we tried were the Open Quotient measures **OQ1** and **OQ2** [13], the Quasi-Open Quotient **QOQ**, and the Speed Quotients **SQ1** and **SQ2**.

For each measure, values every 5ms were found as follows: we calculated the value of each measure in overlapping segments of 32 ms and 64 ms (also stepped every 5ms) using Aparat [14] and then defined the value of a measure at time t to be the mean of its values in all segments containing t.

3.2. Harmonic-Formant Differences

Other common measures of voice quality come from careful analysis of the harmonics and formants of the speech signal, such as the differences **H1−H2** and **H1−A3** [2, 15]. H1 is the amplitude of the first harmonic of a segment of speech, while H2 is the amplitude of the second harmonic. A3 is the amplitude of the largest harmonic in the third formant.

We used the method and Praat script of [15] to calculate harmonics and formants.

3.3. Spectral Summary Measures

The **Spectral Center of Gravity (SCG)** was proposed in [16] as a summary measure for Spectral Balance, and was shown there to correlate with lexical stress in American English. It is higher when there is more energy at higher frequencies. If $|S(f)|$ is the energy at frequency f, then the SCG is $(\int f|S(f)|df)/(\int |S(f)|df)$.

The **Spectral Tilt** of a short segment of speech is defined to be the gradient of the line of best fit to its spectrum between 500 and 4000Hz.

3.4. Band Energy

Band Energy is the energy in each of a collection of frequency bands. This is much easier to calculate than most of the measures previously calculated as no pitch calculation or inverse filtering is required[2]. The energy was measured using the multi-taper spectrogram [17] by considering overlapping 20ms frames of speech stepped every 5ms.

One of the earliest band energy measures suggested for an intonational recognition task was **Spectral Balance** [18], which uses the bands 0-500, 500-1000, 1000-2000 and 2000-4000 Hz. A similar measure, which we denote as **vSN Balance**, using bands 100-300, 300-800, 800-2500, 2500-3500 and 3500-8000 Hz, helps to predict pitch accent and stress in American English [19]. We also used these other sets of bands:

EQ31 has the thirty-one overlapping bands of 250 Hz bandwidth between 0 and 4000Hz: 0-250 Hz, 125-375, 250-500,..., 3750-4000.

EQ15 has fifteen overlapping bands of 500 Hz bandwidth between 0 and 4000Hz : 0-500, 250-750, 500-1000, ..., 3250-3750, 3500-4000.

EQ8 has a subset of bands of EQ15 : 0-500, 500-1000, 1000-1500, ..., 3500-4000.

4. EXPERIMENTS I : VQ MEASURES

The Harmonic-Formant and Glottal Flow features took a particularly long time to compute. Therefore, the experiments reported here only used twenty stories with 1383 syllables. To make up for this, we performed four-fold cross-validation with five stories per fold. We computed performance with varying feature sets; each set consisted of PID68 plus a d-dimensional VQ feature.

For features other than the band energy features, if a syllable had ℓ frames with values $x_1, \ldots, x_\ell$ then the value of the feature for the syllable is a d=4-dimensional vector consisting of the mean and standard deviation of the ℓ values, the midpoint $x_{\lfloor \ell/2 \rfloor}$, and the gradient of the line of best fit.

For band energy features with d bands, we took the value of such a feature for a syllable to be a vector with the mean (over all frames in the syllable's rhyme) of each band.

Table 1 has the results. The best features (which particularly help in recognizing neutral tones) were those based on band energy. This cannot be attributed merely to such features having more dimensions since even Spectral Balance, which has $d = 4$, works better than most non-band-energy features.

Despite the small size of the dataset, there is enough evidence to suggest that band energy features, particularly EQ15, are an appropriate measure of VQ for our purposes.

5. EXPERIMENTS II: BAND ENERGY

We now perform more experiments with EQ15 using a much larger subset of 1159 stories spanning ∼10 hours of speech, with ∼120 000 syllables for training and ∼40 000 for testing.

EQ15 consists of fifteen bands, each of 500Hz in bandwidth. We refer to bands according to their mid-frequency: The first band **B250** covers 0-500 Hz, the second band **B500** covers 250-750 Hz, ..., **B3750** covers 3500-4000 Hz.

Suppose a syllable s has $\ell := \ell_s$ frames in its rhyme. Let x_{in}, for $i = 1, \ldots, \ell$ and $n = 1, \ldots, 15$, be the energy in the n-th band for the i-th frame. For each band n we computed six types of features: the Mean and standard deviation (Stdv)

[2]Preliminary experiments where we used inverse filtering produced worse results; finding a good inverse filter is difficult.

Table 1. Classification performance using a variety of VQ features in addition to the core set PID68 of features based on overall intensity, pitch, and duration. The baseline, using PID68 and no VQ features, is in bold.

	Acc	MeanF	d
EQ15	0.6081	0.5594	15
vSN Balance	0.6066	0.5521	5
EQ8	0.6035	0.5613	8
EQ31	0.6002	0.5585	31
Sp. Tilt	0.5945	0.5318	4
H1−H2	0.5911	0.5195	4
Sp. Balance	0.5907	0.5345	4
AQ	0.5900	0.5214	4
QOQ	0.5892	0.5169	4
H1−A3	0.5870	0.5191	4
ClQ	0.5866	0.5174	4
NAQ	0.5862	0.5194	4
OQ2	0.5862	0.5161	4
OQa	0.5862	0.5155	4
—	**0.5862**	**0.5132**	0
OQ1	0.5858	0.5173	4
SQ1	0.5847	0.5079	4
SCG	0.5840	0.5095	4
SQ2	0.5809	0.5068	4

of $x_{1n}, \ldots, x_{\ell n}$, the Gradient of line to best fit to $x_{1n} \ldots x_{\ell n}$, the Midpoint $x_{\lceil \frac{\ell}{2} \rceil n}$, and the differences MeanMstart $= \mu_n - x_{1n}$ and MeanMmid $= \mu_n - x_{\lceil \frac{\ell}{2} \rceil n}$.

Thus we considered ninety Band Energy features using six types of measurements in fifteen bands. With them only, accuracy was 45.70%, and MeanF 0.4185. When added to PID68, performance was 64.06% and 0.6187 respectively. This is an improvement on using PID68 only, when performance is 60.40% and 0.5400 respectively. Most of the improvement is for neutral tones, for which the F score increases from 0.3447 with PID68 to 0.6175 with the additional band features. All improvements are statistically significant at $p << 0.01$.

6. EXPERIMENTS III : SUBSETS OF BAND ENERGY FEATURES

It is possible that not all 90 features are necessary, so we performed two more sets of experiments: in the first, we performed 15 experiments; in each we used PID68 and all the six types (gradient, meanMstart, etc) associated with one frequency band. In the second, we performed six experiments; in each, we used PID68 and one of the six types for all 15 bands. Detailed results are in [5]; we highlight some here.

While all energy bands contribute to recognition, some are more important. Listing the 15 bands in descending order of classification accuracy when they are added to PID68, we have B500, B750, B1750, B2500, B2250, B2000, B1500, B2750, B250, B1250, B1000, B3250, B3500, B3000, B3750 Energy below 500Hz has often been dismissed as a measure of vocal strength, so it is unsurprising that B250 is one of the less useful bands. On the other hand, B500 is definitely the most useful band, so perhaps it is the energy in frequencies below 250Hz, rather than 500Hz, that is a poor cue for VQ. Frequencies above 3000Hz are not very useful either, though they are still useful; even B3750 provides an increase in accuracy when added to PID68.

Things are clearer when considering types of features: the most important are unquestionably Mean and Mid, followed by MeanMstart and Gradient. At the other end, MeanMend and Stdv are not very useful. If we drop them, i.e. use 60 features instead of 90, we can almost match the performance with 90 bands, with classification accuracy 63.7%, and mean F score 0.6116 (though the difference remains significant).

Table 2. Classification performance using PID68 only.

	Precision	Recall	F
High	0.6089	0.5867	0.5976
Rising	0.5996	0.6789	0.6368
Low	0.5620	0.3822	0.4550
Falling	0.6196	0.7202	0.6662
Neutral	0.5409	0.2529	0.3447
Mean	0.5862	0.5242	0.5400

Table 3. Performance using PID68 and 90 band energy features.

	Precision	Recall	F
High	0.6406	0.6298	0.6351
Rising	0.6327	0.6763	0.6538
Low	0.5965	0.4270	0.4977
Falling	0.6517	0.7323	0.6897
Neutral	0.7111	0.5456	0.6175
Mean	0.6465	0.6022	0.6187

7. CONCLUSIONS

Band Energy features seem far more useful than other possible measures of Voice Quality for Mandarin Tone Recognition. That said, it is possible that such features are more a measure of vocal *strength* than vocal *quality*.

We determined a set of ninety features that when added to a core set of sixty-eight features based on pitch, duration, and overall intensity, improved classification accuracy from 60.4% to 64.1% and the mean F score from 0.540 to 0.619. Improvement is highest for neutral tones, for whom the F score goes from 0.345 to 0.618. This improvement is not at

the cost of other tones; only the F score for Falling tones (the most common class) shows any decrease (and that too only by 0.001).

In fact, it appears that neutral tones can only be recognized using duration and energy; other experiments in [5] failed to recognize any neutral tones using pitch alone. The energy in various frequency bands allows us to characterize neutral tone in a way that isn't possible with pitch.

It remains to be seen if other bands provide better cues than EQ15, and if any can improve the recall (still below 50%) for low tones.

8. ACKNOWLEDGEMENTS

We wish to thank Tae-Jin Yoon for his Praat script to calculate harmonics and formants, Matti Airas and Hannu Pulakka for help with Aparat, Partha Niyogi for multitaper spectral analysis code, and the NSF (Grant 0414919) for funding.

9. REFERENCES

[1] A Belotel-Greni and M Greni, "The creaky voice phonation and the organization of chinese discourse," in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing 2004*, 2004.

[2] Melissa Epstein, *Voice Quality and Prosody in English*, Ph.D. thesis, University of California at Los Angeles, 2002.

[3] Britta Lintfert and Wolfgang Wokurek, "Voice quality dimensions of pitch accents," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005, pp. 2409–2412.

[4] Nick Campbell and Mary E. Beckman, "Accent, stress, and spectral tilt," *The Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 3195–3195, 1997.

[5] Dinoj Surendran, *Analysis and Automatic Recognition of Tones in Mandarin Chinese*, Ph.D. thesis, The University of Chicago, 2007.

[6] Charles L Wayne, "Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2000.

[7] Gina-Anne Levow, "Context in multilingual tone and pitch accent recognition," in *Proceedings of the 9th European Conference of Speech Communication and Technology*, 2005.

[8] Ting-Fan Wu, Chih-Jin Lin, and Ruby C. Weng, "Probability estimates for multi-class classification for pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[9] Sathiya Keerthi and Dennis DeCoste, "A modified finite newton method for fast solution of large scale linear svms," *Journal of Machine Learning Research*, vol. 6, pp. 341–361, 2005.

[10] John Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A.J. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, Eds., 2000, pp. 61–74.

[11] Hannu Pulakka, "Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography," M.S. thesis, Helsinki University of Technology Acoustics Laboratory, 2005.

[12] Paavo Alku and Tom Backstrom, "Normalized amplitude quotient for parametrization of the glottal flow," *Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[13] E Holmberg, R Hillman, and J Perkell, "Glottal airow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 511–529, 1998.

[14] Matti Airas, Hannu Pulakka, Tom Backstrom, and Paavo Alku, "A toolkit for voice inverse filtering and parametrisation," in *Proceedings of Interspeech 2005, http://aparat.sourceforge.net*, Lisbon, Portugal, 2005.

[15] Tae-Jin Yoon, Jennifer Cole, Mark Hasegawa-Johnson, and Chilin Shih, "Acoustic correlates of non-modal phonation in telephone speech," in *Proceedings of the 149th Meeting of the Acoustical Society of America*, 2005.

[16] Rob J J H van Son and J P H van Santen, "Duration and spectral balance of intervocalic consonants: A case for efficient communication," *Speech Communication*, pp. 100–123, 2005.

[17] D B Perceval and A T Walden, *Spectral Analysis for Physical Applications*, Cambridge University Press, Cambridge, U.K., 1993.

[18] Agaath M C Sluijter and Vincent J van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2471–2485, October 1996.

[19] Jan P H van Santen and Xiaochuan Niu, "Prediction and synthesis of prosodic effects on spectral balance of vowels," in *Proceedings of IEEE Workshop on Speech Synthesis*, 2002.