Additional Cues for Mandarin Tone Recognition

Dinoj Surendran and Gina-Anne Levow

May 25, 2006

Abstract

Most cues for Mandarin tone recognition involve pitch, overall intensity and duration. This paper investigates ten other possible cues, and finds one that results in modest, but significant, improvement in classification accuracy on a small speaker-independent corpus of Mandarin news broadcast speech. This cue consists of the energies in the sixteen non-overlapping bands of bandwidth 500Hz from 0 to 8kHz. Most of the improvement is in differentiating the neutral tone from the high, rising, and low tones.

1 Introduction

The automatic recognition of tones in Mandarin is a difficult problem that is being approached from several directions. Approaches include correcting for coarticulation and phrase-level declination, and using better machine learning techniques. This paper is part of a project involving all of the above, but is focused on the more mundane task of finding additional sources of local information for tone recognition.

Pitch is the obvious cue for tone recognition, while (overall) intensity and durational features also play a role. We wish to know if other cues can offer any additional information.

To this end, we fixed a dataset of news broadcast Mandarin speech, a classification algorithm, and a set of core features involving pitch, overall intensity and duration. We then found the change in classification accuracy when the core features were augmented with features from a variety of other cues.

2 Task and Evaluation

We considered 10 files of news broadcast speech from the Mandarin VOA TDT2 corpus. Each file had a single news story read (usually) by a single female speaker. The stories had a total of 1383 syllables. They were automatically labelled and force-aligned at the syllable and phoneme level, and then manually checked. Words for which we could not obtain a good alignment were discarded, but this was less than 1% of the total data. We did *not* discard syllables based on the number of voiced segments found.

The classification task is determining which of five tones each Mandarin syllable has. The tone classes are not balanced; see Table 1 for the number of syllables with each tone. The evaluation metric was 10-fold cross-validation accuracy, with the k-th story forming the test data of the k-th fold.

Table 1: *Distribution of tones in the subset of the Mandarin VOA TDT2 Corpus used in most experiments in this paper.*

High	Rising	Low	Falling	Neutral
307	361	186	453	76

First we performed the task on the syllables using only a core set of features based on pitch, overall intensity, and duration (see Section 4 for details), to get a baseline accuracy. Then, for each extra cue considered, we ran the classification protocol using the core features augmented by the features for that extra cue.

3 Classification Algorithm

The 5-class classification algorithm used was a collection of binary classifiers combined with the 1-vs-1 procedure described by Wu, Lin and Weng [1] to produce probability estimates. The binary classifier was a Support Vector Machine [2] with a Radial Basis Function kernel (RBF SVM) with outputs Platt-scaled [3] to produce quasi-probability outputs.

To deal with class imbalance, training examples were importance-weighted in inverse proportion to the empirical class probability. In other words the importance of all training examples of class j was $1/p_j$, where p_j is the number of training examples of class j divided by the total number of training examples.

The parameters C, γ required for training the RBF SVM were found by 3-fold cross-validation on each fold's training set.

4 Core Features

Our core features were mostly the local features used in [4]. We deliberately excluded features that used context or phrase-level declination correction — we only tested local features for each extra cue (to avoid confounding effects), and it seemed only fair to limit ourselves to local features in the core set.

The temporal features for each syllable were its duration, the duration of its rhyme, and the number of voiced samples. The first two were in seconds, while the last was z-normalized with respect to the corresponding distribution for syllables in the same story. In other words, if the number of voiced samples of the k syllables in a story were $n_1, \ldots, n_k$, then the actual features used were $(n_i - \mu)/\sigma$, where $\mu = \frac{1}{k}\sum_{i=1}^{k} n_i$ and $\sigma^2 = \frac{1}{k-1}\sum_{i=1}^{k}(n_i - \mu)^2$.

Pitch and intensity values were obtained with Praat [5]. We always dealt with the logarithm of pitch, but we will refer to this as pitch for convenience.

Thus each syllable had pitch and intensity values for all its frames. These were used to obtain the features below. We effectively defined a syllable by its rhyme, as in [4], to reduce the influence of syllable-initial consonantal effects.

The pitch of each voiced frame was z-normalized by the distribution of for all voiced frames in its story; likewise for the intensity of each frame. Pitch in unvoiced regions was defined using linear interpolation.

- the pitch contour across the rhyme. Since syllables vary in duration, this was represented by values of 10 equally spaced points along the rhyme.
- the derivative of pitch contour. This was represented by the 9 differences of the preceding feature.
- gradient and intercept of the line of best fit to the pitch contour across the rhyme. These were then z-normalized by the corresponding distribution for all syllables in the same story.
- pitch at the start, middle, and end of the rhyme.
- mean, maximum, minimum of pitch in the rhyme.

Table 2: *Confusion Matrix and per-class accuracy when classifying with a RBF SVM using core features described in section 4. The overall classification accuracy was 65.9%. The confusion matrix is the sum of the confusion matrices for each of the ten folds.*

True	Predicted tone					Acc.
tone	high	rise	low	fall	neut	(%)
high	200	35	8	62	2	65.1
rise	45	270	21	23	2	74.8
low	11	31	92	47	5	49.5
fall	44	25	34	345	5	76.2
neut	4	14	18	36	4	5.3

- All the above features, but for intensity instead of pitch.

5 Additional Cues

Most extra cues we considered have something to do with the 'strength of a syllable' in some sense.

Some make use of spectral analysis; this was a multitaper spectrogram [6] computed with overlapping 20ms frames of speech stepped every 5ms.

Most measures described here were defined for frames rather than syllables. The value of a feature for a syllable was taken to be the mean of the feature values for all frames in the syllable's rhyme. Values were z-normalized by the corresponding distribution for all syllables in a story.

5.1 Voice Quality

Voice Quality (VQ) measures how far speech is from modal speech [7]. It is hard to define perceptually, as listeners disagree on judging modality away from categorical extremes [8], but articulatorily it measures the tension of the vocal folds during speech [9]. It is useful for detecting phrase boundaries in English [7] and Swedish [10]. In vowel-by-vowel analyses, it is useful for detecting pitch accent in German [11] and prominence (narrow focus) in English [12]. It is not a useful cue for detecting stress in Dutch [13].

The literature does not suggest that VQ helps recognize tones, but this has not yet been checked empirically.

Most VQ measures are based on estimating glottal flow during speech and matching it to idealized templates. We computed the **Normalized Amplitude Quotient (NAQ)** [14], **Open Quotient (OQ1)** [15], and the **Speed Quotient (SQ1)** using Aparat [16].

Other methods, such as **Spectral Tilt**, are more indirect. We defined Tilt as the gradient of the line of best fit to the energy spectrum between 500Hz and 4000Hz.

5.2 Band Energy

Band Energy is the energy in each of a collection of specified frequency bands. Sluijter and van Heuven [13] note that increased effort shortens the closing phrase of the glottal pulse, which leads to higher energies above 500Hz. Thus Band Energy cues with bands above 500 Hz should measure vocal effort.

Sluijter and van Heuven go on to show that the bands 0-500, 500-1000, 1000-2000 and 2000-4000 Hz predict stress in Dutch sentences. They call the resulting cue **Spectral Balance**.

There are, of course, infinitely many other choices of bands. van Santen and Niu [17] find that a weighted combination of the energies in bands 100-300, 300-800, 800-2500, 2500-3500 and 3500-8000 Hz correlates with pitch accent and stress in American English. We will refer to this set of bands as **vSN**.

Finally, we use three new sets of bands:

EQ16 also covers frequencies up to 8kHz, but does not do so in a log scale. Instead it has the 16 non-overlapping bands, of 500Hz bandwidth each, i.e. 0-500, 500-1000, 1000-1500, ..., 7000-7500, 7500-8000Hz.

EQ7 has the seven bands of EQ16 between 500 and 4000 Hz.

Bal+EQ7 has the eight bands of EQ16 below 4 kHz plus the bands 1-2 kHz and 2-4 kHz. It is also the union of EQ7 with the bands for Spectral Balance, hence our name for it.

5.3 Spectral Center of Gravity

If $A(x)$ is the energy at frequency x, then the **Spectral Center of Gravity (SCoG)** is $(\int xA(x)dx)/(\int A(x)dx)$. It was proposed in [18] as a summary measure for Spectral Balance, and was shown there to correlate with lexical stress in American English.

6 Results

Baseline classification accuracy with the core features was 65.9%. However, this varied hugely across tones, from 5.3% (neutral) to 76.2% (falling), as shown in Table 2. This is partly due to the limited training data, but is likely more due to the core features not capturing anything that can separate neutral tones from the rest.

Table 3 shows results for all extra cues. The best cue, EQ16, results in classification accuracy 69.2%. All cues that improved on the baseline were Band Energy cues. All voice quality cues misled the classification algorithm (but probably not significantly).

Most improvement for EQ16 is in recognizing the neutral tone. Though its accuracy/recall is still low, at 38.2%, it is much higher than the baseline 5.3%; see the bottom rows of Tables 2 and 4. EQ16 clearly helps separate neutral toned syllables from high, rising, and low toned syllables — but not from falling toned syllables.

7 Discussion

Only measures of Band Energy offered any improvement to the core features, with most improvement for the neutral tone.

This could be because band energy is correlated with stress (even though the notions of stress in Mandarin and the other languages mentioned previously are not identical) and neutral tones are never on stressed syllables.

On the other hand, the only band energy feature that did not result in an improvement to the baseline was Spectral Balance, precisely the feature that Sluijter and van Heuven [13] found predicted stress in Dutch. However, it did result in a small improvement in recognizing the neutral tone — it was just that this was at the expense of classification accuracy on other tones.

Table 3: *Classification accuracies, overall and per-class, using a variety of additional cues. Balance, vSN, EQ16, EQ7, and Balance+EQ7 are Band Energy cues, while Tilt, NAQ, SQ1 and OQ1 are Voice Quality measures. The baseline uses no additional features.*

Extra Features	Accu-racy	Per-tone Accuracy high	rise	low	fall	neut
EQ16	69.3	69.4	76.7	46.2	77.9	38.2
Bal+EQ7	68.5	68.7	76.2	46.2	76.4	38.2
EQ7	68.3	67.1	76.7	47.3	75.9	38.2
vSN	67.6	68.7	75.4	45.2	77.5	22.4
None	65.9	65.1	74.8	49.5	76.2	5.3
SQ1	65.7	64.8	73.4	46.2	77.9	6.6
Balance	65.5	66.1	73.1	42.5	76.8	15.8
SCoG	65.5	64.2	72.6	46.2	77.9	10.5
Tilt	65.4	64.5	72.0	46.2	77.5	11.8
NAQ	64.8	61.9	74.5	43.0	78.1	3.9
OQ1	64.2	62.8	73.1	45.7	75.5	5.3

7.1 Significance Testing

An important question is whether the difference in classification accuracy, which is just over 3% in absolute terms, is statistically significant ($p < 0.05$). We can easily get significance using a 10-fold cross-validated paired t test, but that test often says there is a significant error rates when there is not, because the overlap of the training sets between folds invalidates independence assumptions.

Therefore, we ran the conservative 5x2CV test [19]. This is normally used to see if there is a significant difference in error rates when two algorithms run on the same features, but can also be used when changing features instead of algorithms. It consists of running 2-fold cross validation on the full training set 5 times, and seeing if the difference in error rates is consistently large enough. The 2-fold cross validation ensures that there is no overlap in training examples on each run.

Using just the ten stories, we were unable to get significance. Although there was a reduction in error rates when using EQ16 as additional features for each of the 10 training-test splits tested, it was not large enough to be significant ($p = 0.101$).

However, we had more data available, from ten more stories that we had excluded from our previous experiments to prevent overtraining during feature selection. Repeating the 5x2CV test with the 2672 syllables in the 20 stories, we found significant improvement ($p = 0.043$).

If one does not mind breaking various statistical assumptions, one can also apply the 5x2CV significance test to per-tone classification accuracy. In this case we found significant improvement ($p = 0.040$) in recognition for the neutral tone only. Changes for the other three tones, including the loss of accuracy for low tone, were not significant.

7.2 Importance of bands

It would be useful to know the relative importance of the 16 bands in EQ16. Unfortunately, getting this when using a RBF SVM, especially in a multiclass setting, is difficult[1].

[1]Such methods exist, but we did not have the time to implement them.

Table 4: *Confusion Matrix and per-class accuracy when classifying with a RBF SVM using core features and band energy feature EQ16 described in section 5.2. Overall accuracy was 69.2%.*

True	Predicted tone					Acc.
tone	high	rise	low	fall	neut	(%)
high	213	34	7	49	4	69.4
rise	34	277	21	26	3	76.7
low	14	33	86	48	5	46.2
fall	40	26	28	353	6	77.9
neut	4	8	6	29	29	38.2

Thus we found instead the importance of each band when using a less powerful, but more interpretable, binary classifier: a linear SVM. Unsurprisingly, despite choosing its parameter to minimize *test* error (as we were interested in weights, not accuracy) the linear SVM did not work as well as the RBF SVM. Its accuracy was 60.5% and 62.5% without and with EQ16 respectively.

The trained linear SVM for each binary classification problems produces a weight w for each dimension. The number of dimensions here is 73, as each syllable is represented by 57 core features and the energy in 16 bands. The absolute value $|w|$ of the weight indicates the importance of the dimension.

Recall that each 5-class classification problem is converted to 10 binary classification problems in a 1-vs-1 strategy. One way of expressing the importance of a dimension is to consider the mean value of $|w|$ across all binary problems; another is to consider the maximum value. We shall refer to these two measures as $mean\ |w|$ and $max\ |w|$ respectively. The two measures are complementary; the first rewards dimensions that are moderately useful across all binary problems while the second rewards dimensions that are extremely useful for some binary problem.

Table 5 shows the values of these two importance measures, averaged across each of the 10 folds, for each band. The bands with the most information are 1500-2000, 2500-3000, 3500-4000, 3000-3500, and 500-1000. However, all bands are in the top half of all 73 features when ranked by either importance measure.

While keeping in mind all caveats about features useful for linear SVMs not necessarily being those useful for RBF SVMs, it would seem that the reason the bands of Spectral Balance are not useful here is that they are too coarse-grained in the area of frequency space where the information is. They use two bands to cover 1-4 kHz while EQ7 uses six bands there.

8 Conclusions

The Band Energy feature EQ16, with intensities in the 16 non-overlapping 500Hz-wide bands from 0 to 8kHz, is a useful feature for the recognition of neutral tone in Mandarin. Most, but not all, information is below 4 kHz.

Voice Quality does not appear to be useful for this task. On the other hand, this could be due to the way we computed Voice Quality. Here we did not inverse filter the data, but we did do so in unreported preliminary experiments and always obtained even worse results. It is possible that we did not inverse filter correctly, and since one would expect voice quality to help with the recognition of third tone, we will redo those experiments with alternative inverse filtering methods.

There are several open questions. The first is understanding why the choice of bands in EQ16 works,

Table 5: *Importance of bands in EQ16 when classified with a linear support vector machine. The rank is with respect to all 73 features used (57 core features and 16 bands), with 1 being the most important feature.* $Mean\ |w|$ *and* $max\ |w|$ *are both possible measures of importance; see Section 7.2 for details.*

band	$mean\ \|w\|$		$max\ \|w\|$	
	value	rank	value	rank
0 - 500	0.35	18	0.98	10
500 - 1000	0.48	8	0.81	17
1000 - 1500	0.26	29	0.61	28
1500 - 2000	1.03	1	2.51	1
2000 - 2500	0.36	16	0.97	11
2500 - 3000	0.63	3	1.86	2
3000 - 3500	0.50	7	1.17	8
3500 - 4000	0.58	4	1.45	4
4000 - 4500	0.34	20	0.68	21
4500 - 5000	0.26	28	0.60	29
5000 - 5500	0.43	10	0.84	15
5500 - 6000	0.38	14	0.64	25
6000 - 6500	0.40	13	0.85	14
6500 - 7000	0.28	27	0.68	23
7000 - 7500	0.23	34	0.57	33
7500 - 8000	0.23	35	0.61	27

especially as it is not logarithmic in frequency. This will help design better bands.

Second, is EQ16 (just) capturing stress? After all, Band Energy features predict stress in other languages while Voice Quality features do not, and the former clearly worked better here. Of course, whether stress in Mandarin has any relation to stress in English, German or Dutch is another matter entirely. It would be useful to repeat this experiment on a corpus of Mandarin syllables where stress is known [20].

9 Acknowledgements

We would like to thank Chih-Jen Lin, Chih-Chung Chang and their collaborators for writing LIBSVM [21] and its associated tools, which we used for all the SVM experiments. Thanks also to Partha Niyogi for the original multitaper spectral analysis code, and to Matti Airas, Tom Backstrom and Hannu Pulakka for help with using Aparat.

References

[1] 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.

[2] Corinna Cortes and Vladimir Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[3] 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.

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

[5] Paul Boersma and David Weenink, "Praat: doing phonetics by computer," *http://www.praat.org*, 2005.

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

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

[8] Jody Kreiman and B Gerratt, "The perceptual structure of pathologic voice quality," *Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1787–1795, 1996.

[9] 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.

[10] Christer Gobl, "Voice source dynamics in connected speech," Tech. Rep., Royal Institute of Technology, Stockholm, 1988.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Thomas G. Dietterich, "Approximate statistical test for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[20] Min Chu, Yunjia Wang, and Lin He, "Labeling stress in continuous mandarin speech perceptually," in *Proceedings of the 15th International Congress of Phonetic Sciences*, 2003.

[21] Chih-Chung Chang and Chih-Jin Lin, "Libsvm : a library for support vector machines," *Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm*, 2001.