

Tone Recognition in Mandarin using Focus

Dinoj Surendran, Gina-Anne Levow, Yi Xu

Interspeech 2005

dinoj@cs.uchicago.edu
levow@cs.uchicago.edu
yi@phonetics.ucl.ac.uk

You might find this poster interesting if...

- ▶ You want to know how to classify tones better
 - ▶ We reduce the error rate of a linear SVM from 15.1% to 9.2% using focus without ever knowing which syllables are focused (even in training)
- ▶ You care about how focus affects tones.
- ▶ You want to find out about a new predictor for focus in a tonal language.
- ▶ You want to see an example of how to predict A using B when B is unknown.
- ▶ You want to give the poor sod standing here something to do

Introduction

- ▶ Low-level tone recognition is important.
- ▶ Wang & Seneff (2000) studied effect of coarticulation, etc (but not focus) on tone - we fill gap in their study. (Unlike them, we have a focus-marked dataset.)
- ▶ Rough definition: a word in a phrase is focused if it is emphasized. (Focus is sometimes called 'prominence'.)
- ▶ We use focus to mean 'narrow focus' only. (What we call 'sentences with no focus' others call 'sentences with broad focus'.)
- ▶ We propose a no-training-required predictor of focus, and use predicted focus to improve tone classification accuracy.

- ▶ XTF99 : 11520 syllables with tones 1-4 in 3840 phrases by 8 native Mandarin speakers
- ▶ Previously analyzed in Xu's 'Effects of tone and focus on the formation and alignment of f0 contours', J. Phonetics (1999)
- ▶ Clean lab speech, manually segmented
- ▶ Tone and **Focus marked**
- ▶ All words are monosyllabic (for our purposes).
- ▶ Each syllable normalized to 20 samples
- ▶ Available online so you can run your own experiments

<http://people.cs.uchicago.edu/~dinoj/projects/tonefocus>

Classifying by Groups

- ▶ Any classification algorithm (we use a linear SVM) is given tone-labelled syllables and learns a function $R : \{\text{syllables}\} \rightarrow \{\text{high, rising, low, falling}\}$.
- ▶ Applying LinSVM to all syllables results in classification error rate of 15.1%.
- ▶ Xu (1997) noted that post-focus syllables have lower pitch range than in-focus syllables and pre-focus syllables.
- ▶ Perhaps we can improve performance by finding different rules for syllables with different focus conditions:
 - ▶ no-focus syllables : those in sentences with no focus.
 - ▶ in-focus syllables: those in a focused word of a sentence.
 - ▶ pre-focus syllables : those before in-focus syllables.
 - ▶ post-focus syllables: those after in-focus syllables.
- ▶ Using LinSVM to create a different decision rule for each group above results in an error rate of 7.9%.

Classifying by Groups of Focus Conditions

- ▶ Given partition G of $\{\text{syllables}\}$ & classification algorithm C , C_G is combination of separate classifiers for each class in G .
- ▶ e.g. error rate for $LinSVM_{\{\{no,pre,in,post\}\}}$ is 15.1%
- ▶ e.g. error rate for $LinSVM_{\{\{no\},\{pre\},\{in\},\{post\}\}}$ is 7.9%. Error rates for individual groups are below:

no-focus	pre-focus	in-focus	post-focus	combined
7.0	7.4	0.7	16.7	7.9

What this table says:

- ▶ In-focus syllables are very easy to recognize.
- ▶ Pre-focus and no-focus syllables are recognizable with similar, average difficulty.
- ▶ Post-focus syllables are very hard to recognize (focus has asymmetric effects on pre- and post-focus syllables).

Classifying by Groups of Focus Conditions II

Partition G of focus conditions	Error Rate of LinSVM $_G$
{no, pre, in, post}	15.1
{no, pre, post}, {in}	14.6
{no, post}, {in}, {pre}	11.2
{no, pre}, {in}, {post}	8.2
{no}, {pre}, {in}, {post}	7.9

- ▶ This table suggests that no- and pre-focus syllables can be grouped together with little loss of accuracy. This means the two kinds of syllables 'behave' similarly.
- ▶ Computational Bonus: if we can lump together no-focus and pre-focus syllables then we never have to make a decision as to whether phrase has focused syllables or not.

An algorithm for splitting by focus condition

- ▶ We want to use $\text{LinSVM}_{\mathcal{P}}$, where $\mathcal{P} = \{\{\text{no}, \text{pre}\}, \{\text{in}\}, \{\text{post}\}\}$, but do not know the focus condition of any syllable.
- ▶ **Hypothesis:** Since in-focus syllables are easiest to recognize, we predict that the **syllables recognized with most confidence in a sentence are focused**.
- ▶ Note: we need not decide which syllables are no-focus.

Phase 1 : produce an estimate $\hat{\mathcal{P}}$ of \mathcal{P} .

- ▶ Train multiclass LinSVM (on all syllables) to produce posterior probability estimates (using Wu, Lin & Weng, JMLR 2004) $P(y|x)$ of the tone y of any syllable x .
- ▶ Given syllables in phrase x_1, \dots, x_n , predict that the focused syllable is x_j if $j = \text{argmax}_{i=1, \dots, n} \max_y P(y|x_i)$. This labels other syllables in the phrase as pre- or post-focus.

Phase 2: train $\text{LinSVM}_{\hat{\mathcal{P}}}$

Results : focus prediction (estimating \mathcal{P})

- ▶ All our phrases have 3 syllables: a phrase has n -focus if the n -th syllable is focused (for $n > 0$) and 0-focus if no syllable is focused.
- ▶ Error Rate predicting n -focus of phrase = 36.3%

	1-focus	2-focus	0- or 3-focus
0-focus	127	300	533
1-focus	646	81	233
2-focus	51	727	182
3-focus	99	322	539

- ▶ Error Rate predicting focus condition of syllable = 32.8%

	no or pre-focus	in-focus	post-focus
no-focus	1366	960	554
pre-focus	2309	472	99
in-focus	496	1912	472
post-focus	233	496	2151

Results : tone prediction (using the estimated $\hat{\mathcal{P}}$)

- ▶ Error Rate when predicting using $\text{LinSVM}_{\hat{\mathcal{P}}}$ is 9.1%

Combined (error rate = 9.1%)				predicted pre-focus (error rate = 10.1%)			
3889	131	54	86	1300	57	21	45
306	1877	47	10	104	1047	30	7
25	56	2741	58	0	40	656	36
217	18	44	1961	71	14	19	957
predicted in-focus (error rate = 2.4%)				predicted post-focus (error rate = 15.7%)			
885	9	13	7	1704	65	20	34
15	716	6	0	187	114	11	3
1	7	1254	8	24	9	831	14
16	0	12	891	130	4	13	113

- ▶ Compares with 15.1% baseline and 8.2% with true $\text{LinSVM}_{\mathcal{P}}$.

Conclusions

- ▶ It is possible to use focus to improve tone recognition even when focus labels are not available — but remains to be seen if this holds for real-world data (future work)
- ▶ In-focus syllables are very easy to recognize. The confidence of predicting a syllable's tone is a good predictor of whether the syllable is focused.
 - ▶ In our experiments, it is at least as good as explicitly training a classifier on lots of focus-labelled examples.
- ▶ Pre-focus syllables behave like no-focus syllables.
- ▶ Focus affects post-focus sylls differently from pre-focus sylls.
 - ▶ Xu (97): post-focus syllables have lower pitch range & mean.
 - ▶ e.g. rising and falling post-focus syllables are more likely to be labelled as high — the pitch in high post-focus syllables tends to approach the mean pitch value (usually it approaches an above-average pitch value) from above or below depending on the preceding tone.