Chinese syllable and tone discrimination

Isac Arnekvist
isacar@kth.se

Abstract

A convolutional neural network was trained to discriminate between 46 classes of chinese syllables (Mandarin) and the tones 1-4. Convolutions were done on fixed length features both in time and frequency domain. Four female and three male speakers were used for training. The network was found to classify previously unseen utterances from seen speakers really well, but much poorer for new speakers. It is reasonable that really good results could be acheived with this architecture with more data, but to apply it to continuous speech recognition might be problematic due the setup relying on fixed sequence length.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Table 1: The pinyin (chinese phonetic writing) of all the recorded utterances. In total 46 classes.

t	zh	q	r	z	c	x	j	sh	s	ch
ta	zha	qi	ri	zi	cao	xi	ji	sha	sa	cha
ti	zhi	qv	re	ze	ci	xv	jie	shi	si	chi
te	zhe		ru	zu	ce	xia	jv	she	se	che
tou	zhuo			zuo	cu			shuo	suo	chou
tuo	zhu				cuo			shu	su	chu
tu										

1 Introduction

1.1 Convolutional neural networks

Convolutional neural networks (CNNs) have received a lot of popularity due to their recent advancements. One of the more famous breakthroughs was in the ImageNet contest with an architecture called Alexnet (Krizhevsky et al., 2012). One of the strengths is argued to be shared weights, that the same parameters can be used to detect a feature non-regarding of where in a picture the feature exists. Convolutions have been used successfully in speech for convolving over the frequency domain, which is argued to lower the sensitivity for speaker variation, especially if pooling is used (Abdel-Hamid et al., 2014). This insensitivity of shift in location could be argued to be of interest for the time domain in speech as well since phonemes can have different length and therefore be positioned different in time between different utterances of the same word. Of course the phonemes have to be ordered correctly, but this holds for images as well, e. g. a cat's eyes are above its nose etc and is something still handled by CNNs. Position of features are somewhat kept through the network even though the weights are shared.

1.2 Chinese pronounciation

The chinese spoken language have fricative sounds that often are considered hard to differentiate between for people in the western world. Also, tones are of importance in several Asian languages, including Chinese, in contrast to European languages. It would be interesting to see whether CNNs could be trained to recognize these classes.

2 Method

2.1 Labels/Classes

This project consists of two classification problems, one is what tone the utterance was said with, the other is classification of the phonetic part. The classes of the later part are in total 46 and are listed in table 1. There are in total 4 tones in Mandarin, excluding a fifth tone called "neutral". The neutral tone is omitted in this experiment. A schematic of tones 1-4 are shown in fig. 1.

2.2 Feature representation

The features consists of the outputs from 40 filterbanks spaced accordingly to the the mel frequency scale. Regarding the time axis, all recorded utterances were reshaped to have the same length. Those recordings that needed to be extended were so by repeating start- and/or end frames. This gave an oppurtunity to randomly choose how much to extend either end. This shifting of the input data is called data augmentation or jittering. An example of this on data from this project is shown in fig. 2. Also, gaussian noise were added on top of the features randomly before each epoch to try diminish risk of overfitting.

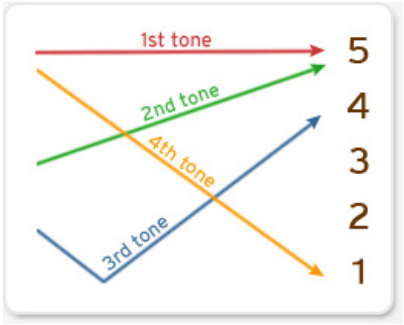

Figure 1: Schematic representation of the four tones. Vertical axis represents pitch and horizontal axis represents time.

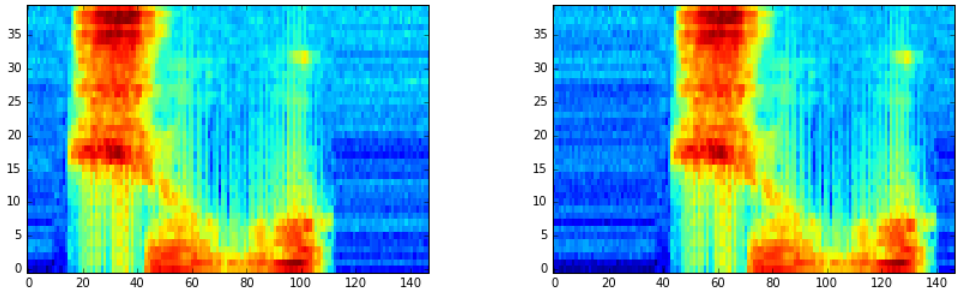

Figure 2: Time/frequency plots showing shifting in time dimension and added noise.

2.3 Training data

All syllables were recorded 4 times, one time for each tone. When first starting this experiment, there were recordings from three speakers. These were permutated and divided into three sets. These results showed generalization with respect to the same speakers, but then a discrepancy between that accuracy and accury with respect to new speakers were noticed, so the experiment setup was changed.

With recording here, one run through of all the classes is meant. Training set consists of three male speakers and 4 female speakers. Most speakers are chinese mother tongue speakers, some have southern chinese accents, but were instructed to try to speak as "standard" as possible. Some speakers made several recordings. Validation set is one male speaker. Test set is one male and one female speaker. Validation set and test set are only mothern tongue speakers or fluent speakers with (at least subjectively) very accurate standard chinese pronounciation. The same mic, a regular iPhone headset, was used for recording. Different rooms and environments were used with a varying degree of surrounding sound and ambience. Utterances were split and labelled using a script that used silences in between utterances as guide where to split.

2.4 Network layout

The layout for the 46-class problem is shown in fig. 3. Earlier experiments have used kernels that move in the 40 dimensional frequency domain with width three, whilst this report experiments with kernel width 38 in the frequency domain, stride 1 and width 6 in the time domain. The same layout was used for tone classification, but the last layer instead was changed to have 4 outputs. All weights excluding biases in the network were initialized randomly from $\mathcal{N}(0, 0.1)$. Bias weights were initialized to zero.

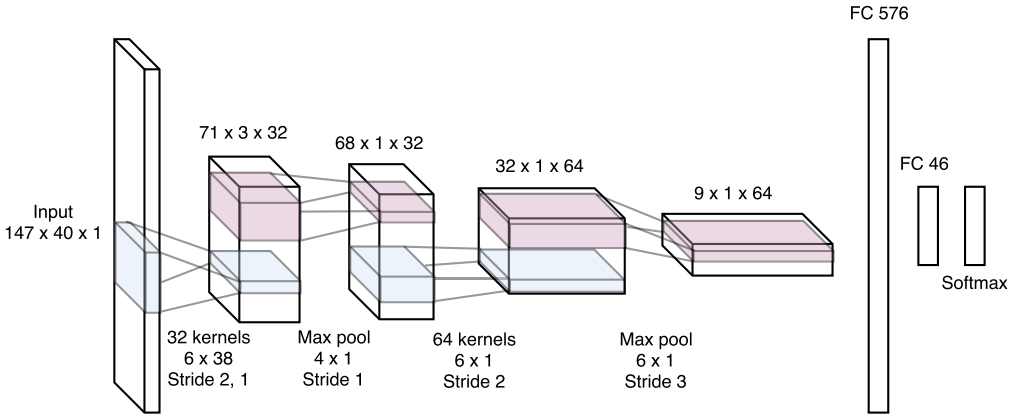

Figure 3: Layout used for syllable recognition. Input consists of 147 time frames and 40 filterbank features per frame. Kernels are "moving" both in frequency and in the time dimension.

Table 2: Test accuracy scores on final models, tones were only evaluated for new speakers in test set

	CNN	SVC(Linear)	SVC(RBF)
Same speaker (syllables)	0.993	0.278	0.239
New speaker (syllables)	0.607	0.337	0.242
New speaker (tones)	0.672	0.595	0.611

2.5 Training

Training was done on the training set and at the same time monitoring the loss on the validation set until convergence or signs of overfitting. Training was done in batches of size 256. The loss was given by the following equation, where $\mathcal{L}_{ce}$ is cross entropy loss, and W are all the weights:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda ||W||^2 \tag{1}$$

The coefficient λ was initially set to 0.001 and after some amount of epochs set to zero. This was informally due to that some experiments showed (no ref. at the moment, sorry) that after some epochs with initial regularization, overfitting becomes less of a problem.

3 Results

3.1 Loss function evolution

Following the loss over time for both the training and the validation set, one notices differences in the rate which they change depending on if validation set includes only same speakers as in training set or unseen speakers (see fig. 4).

3.2 Metrics

Training and validation loss converged without signs of overfitting both in the case of syllables and tones. Results on test set (with same speakers as in training and validation set) had higher scores than for the test set with new speakers. Results are compared with SVM/SVC with C parameter set to 1.0 (penalty coefficient) and using linear and RBF kernels. No hyperparameter optimization was done for the SVCs.

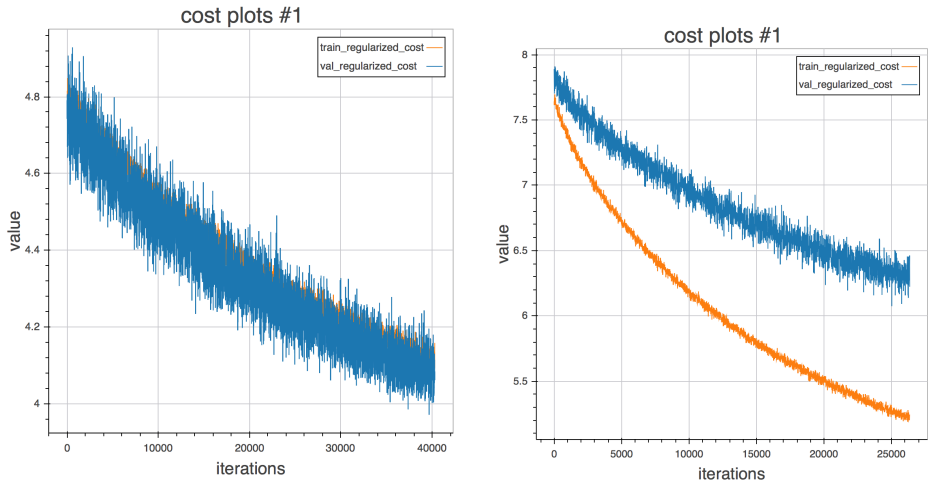

Figure 4: Evolution of the loss on training and validation set. The left one is where validation set has same speakers but new utterances. The traing cost (orange) is hidden under the validation cost (blue). Right plot is when having new speakers in the validation set, here the training set loss diminished in a faster pace.

3.3 Results in more detail

The full confusion matrix for the syllables can be seen in fig. 5. The most common confusions, or mistakes, the network did in order of how many they were are listed in table 3.

Table 3: Most common classification mistakes of this report's network

Correct	Guess
ji	zi
shu	chu
qi	ci
ju	zhi
ti	ji

4 Discussion & Conclusions

4.1 Same speakers in training and validation set

It was interesting to see that the model could generalize so well to unseen data from the same speakers, but so much poorer for a new speaker even though there were data from both male and female speakers. This might be to the fact the kernels were much wider in this experiment in frequency domain compared to the study by Abdel-Hamid et al. (2014). Informally though, the experiments that I did with smaller kernels didn't seem to work very well. This could be due to other factors in the layout of the network that was not taken into account. It seems reasonable that when having smaller kernels, it becomes more important to keep more of location structure until the fully connected part compared to this experiments layout. In this experiment, the frequency domain is simply squashed to one dimension, only keeping dimension width in the time domain.

4.2 Expectation of the performance

The reason for chosing this subset of chinese syllables was due to the fricative part being hard to distinguish for me as a beginner chinese student and many others with me. The syllables that the network performed worst on were in most cases what I think are the hardest ones. What was not expected to be so hard were the sounding parts, the vowel component, of the

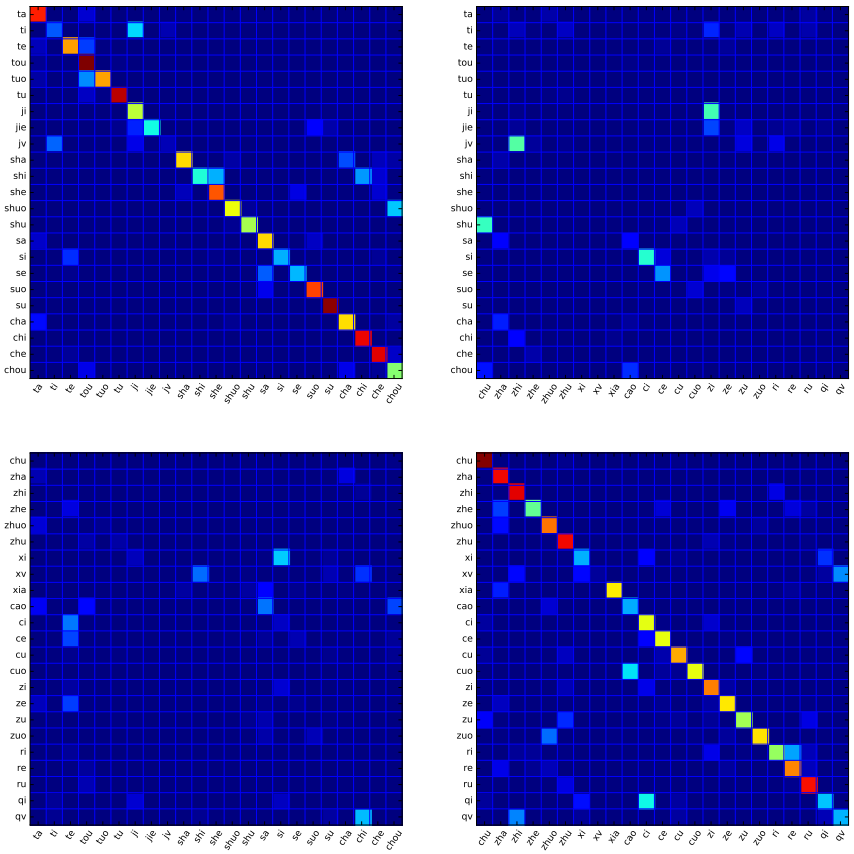

Figure 5: Confusion matrix of the 46 syllables on test set with new speakers, male and female

syllable. This might be though that the vowels are more similar to Swedish pronounciation than Chinese fricative sounds are, but to the network, these are previously unseen of course.

4.3 Implications

It is reasonable to believe that really good results, in a more generalizable sense, can be achived with a much larger dataset. You also understand that speaker normalization techniques might be motivated, although I am not sure how well they work in practice. Then of course, the goal often is to have systems for continuous speech recognition and how to apply this technique to sequences of different length and also containing several words is unknown and might be problematic.

References

O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545, October 2014. URL `http://research.microsoft.com/apps/pubs/default.aspx?id=230894`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`.