Using CNN to Predict if Native Mandarin Listeners Can Correctly Identify the Tone of Target Characters in Noisy Environment

Mengxuan Zhao / Computational Linguistics

Abstract

The significant area are decisive location points in time-frequency spectra for a listener to correctly identify the phonetic information of the target. By adopting the method of "bubble noise" (Mandel et al., 2016), it is proved that such area in time-frequency spectra exists for Mandarin tone identification, and the shape of the area varies for different tones. I am curious about if machines are able to learn the patterns and predict if human listeners can correctly identify the tones, too. I construct four similar convolutional neural network classifiers, each for one tone. All classifiers expect for tone 3 beat the baseline and result in satisfying performance. This paper introduces the background of current task, explains the architecture of the convolutional neural network, demonstrates the classification results, and analyze the difficulties of tone 3 prediction.

1 Introduction

Both human beings and automatic speech recognition systems are reported to be good at identifying speech in noisy environment (Festen and Plomp, 1990; Scharenborg, 2007). Mandel et al. (2016) raised the "bubble noise" method, and located the significant area, with which native English speakers are able to correctly identify six English consonants. Choi (2018) studied the perception of five Korean coronal fricatives and affricates in both native listeners and L2 learners with "bubble noise" mechanism, and compared the difference between the two groups. Both used a linear SVM model to analyze their experiment results, predicting whether a particular mixture of speech and noise is correctly identified by a human listener.

While a number of acoustic cues, such as fundamental frequency (F0), envelope, length etc., play an important role in the perception of Mandarin tones (Liu and Samuel, 2004; Shen and Lin, 1991), the significant area also releases considerable information of intelligibility. The significant area is always positively correlated with the correct identification (figure 1). In pilot study, I generated 4,000 sound files with 5 participants, each is a mixture of a sentence in Mandarin and random "bubble noise", and labeled with the tone perceived by the listener or "don't know" if the listener cannot identify anything. Significant area is found for each of the four tones, however the frequency range and the temporal range of the significant area are different for each tone. The difference is explicable with phonetic reasons. The existence of the significant area justifies that the intelligibility of a speech snippet is predictable. The task of this paper is to build a convolutional neural network model to predict if a particular noisy speech can be correctly identified.

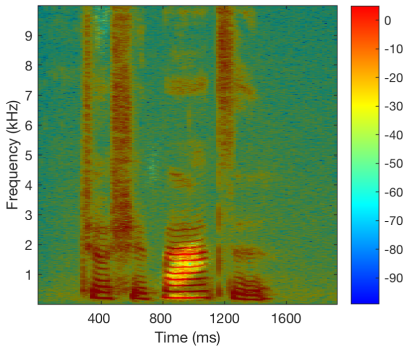

Figure 1: Significant area (highlight)

2 Data

The stimuli for human participants are 2-second speech of a Chinese sentence, combined with white noise with random "bubbles" dig out. The input sentence "這 是 ma 字", meaning "this is the character 'ma'", has four variations, each with a unique tone for the target character "ma". The

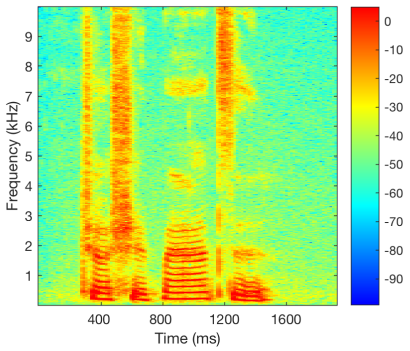

Figure 2: Original pure speech

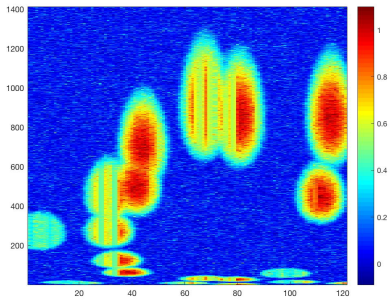

Figure 3: Feature matrix

mixture sound files are generated by a MATLAB based system, and are played to human participants, requiring the participants to choose the tone they hear, or "don't know" if they have absolutely no clue to judge. The mixture sound files, as well as the human labels are used as the database of the CNN classifier. Based on the results of experiments, the significant area of different tones varies. As such, it is sensitive to group all data according to the tone, and predict separately. Figure 2 is one of the original sentences, and figure 3 is the feature matrix derived through Fast Fourier Transform from the combination of figure 2 and random "bubble noise". The size of the feature matrices is $1,412 \times 121$. 1,000 mixtures are used for each tone. The train/dev/test splitting ratio is 900:50:50. In this task binary labels are applied. "1" indicates that the speech snippet is correctly identified and "0" indicates not.

3 Experiment

3.1 Convolutional Neural Network

In order to predict if a specific speech snippet is intelligible, I build a convolutional neural network classifier. Two convolutional layer sets are constructed. In each set, there are two convolutional layers, each with a following non-linear activation layer, and a max pooling layer at the end. In the first set, convolutional filters are 15×15 and 11×11 respectively, enlarging the feature matrices from single channel to 8 channels, and a ReLU function (formula 1) follows each convolution. The max pooling layer, with a kernel size and a stride of 16×9, down-samples the feature matrix to 88×13. In the second set, convolutional filters are 7×7 and 3×3, enlarging the matrices to 32 channels, and tanh function (formula 2) is used as the activation function. The max pooling once more down-samples the matrices with a filter size and a stride of 2×1. The full connected layer takes the flattened matrices of $44 \times 13 \times 32$ channels and maps them to 2-dimension results. The index of the maximum values is taken as the binary prediction results (formula 3).

	Baseline	Precision	Recall	F1 score
tone1 dev	0.68	0.691	0.711	0.696
tone1 test	0.72	0.808	0.867	0.821
tone2 dev	0.68	0.8	0.721	0.740
tone2 test	0.58	0.704	0.669	0.670
tone3 dev	0.7	**0.588**	**0.562**	**0.561**
tone3 test	0.62	**0.569**	**0.565**	**0.566**
tone4 dev	0.64	0.783	0.783	0.783
tone4 test	0.48	0.685	0.667	0.653

Table 1: Experiment results

$$ReLU(x) = max(0, x) \tag{1}$$

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2}$$

$$\overline{y} = argmax(x_1, x_2) \tag{3}$$

I choose cross entropy as the loss function (formula 4), and Adam optimizer to backward optimize the neural network. After the whole convolutional process and after the linear function, dropout is applied, both with 0.4 p value. Other related parameters include batch size = 50, learning rate = 0.0001, number of epochs = 30.

$$loss(x, class) = x[class] + log(\Sigma exp(x[j])) \tag{4}$$

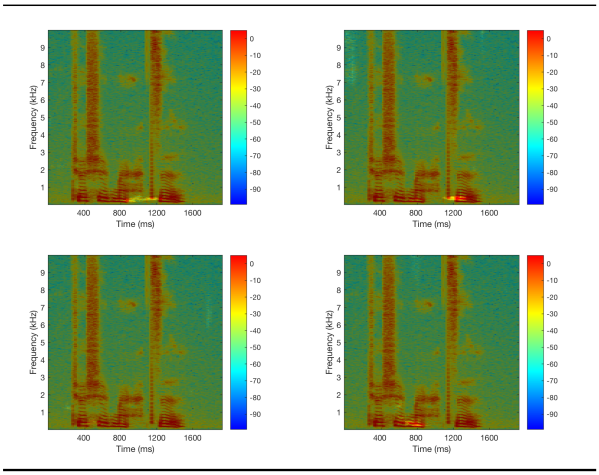

Table 2: Significant area for tone 3

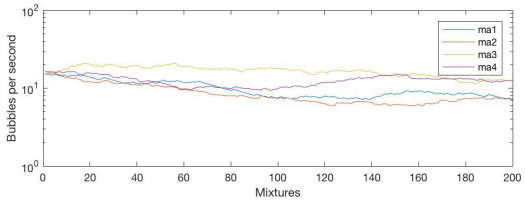

Figure 4: The change of bubble numbers for each tone

3.2 Results and Analysis

Table 1 shows that the CNN models beat the baseline except for the one with tone 3. That corresponds to the results of experiments, where very vague significant area was found for tone 3, and it remarkably varies from participant to participant (table 2). Tone 3 is also the most difficult one to identify for almost all participants. Figure 4 shows the result of one participant, and it is clearly seen that the number of bubbles per second for tone 3 is higher than the other three tones for the majority of time. The "bubbles" are areas where noise is removed and the original speech is revealed. More bubbles mean that the listener requires comparatively more speech information and less noise interference

4 Some Phonetic Facts of Mandarin Tone

The majority of words in Mandarin Chinese are monosyllabic or bi-syllabic. The fact of simple phonetic structures requires elements more than consonants and vowels to deal with the large number of homophones, which results as tones. There are 4 lexical tones in Mandarin[1]. The primary cue to lexical tone identity is the F0 contour, with each tone associated with a particular F0 pattern(Liu and Samuel, 2004). It is clearly seen from figure 1 and table 2 that listeners use F0 information to identify the tone. Chao (1968) marked the pitch values with a 1-5 pitch scale: "55" for tone 1, "35" for tone 2, "214" for tone 3, and "51" for tone 4, where 5 indicates the highest pitch and 1 the lowest.This notation system is a widely used in academic research and pedagogy. However, in reality, tone 2 does not go directly up after onset. Similarly to tone 3, tone 2 also has an initial descending part. Tone 2 and tone 3 have very similar F0 contour shape (Yang, 2010). Liu and Samuel (2004) found that people's performance of distinguishing tone 2 and tone 3 is quite good even when the F0 is neutralized with white noise, which indicates that listeners can use secondary cues when the primary cue is unavailable. Shen and Lin (1991) argued that the ΔF0 from the onset and the turning point, as well as the ratio of timing before and after the turning point are critical for listeners to distinguish the two tones. Both (Liu and Samuel, 2004) and (Shen and Lin, 1991) conducted their experiments on monosyllables. Huang and Holt (2009) argued that the higher context frequency tend to shift perception toward lower frequency targets and vice versa, and the mean frequency of the targets plays as the key in how the spectral energy variations influence the tone perception.

Tone 3's realization is highly context-dependent. In the sentence of my experiment, where the tone 3 "ma" is preceded with a tone 4 character, and followed with a tone 4 character, only the first half of the tone 3 "ma" is produced, in the form of a low falling tone, and the pitch value can be marked as "31"(Yang, 2010). In other words, a rapid pause appears only in the sentence with tone 3 target character, but not the all of the other three stimuli, and it can be seen from table2 that some listeners are using the pause as a clue to identify tone 3.

There is mutual influence between the tone 3 and the context. While the tone of the target character is influenced by the preceding character, it influence the pitch of the following character as well. Xu (1997) claimed that F0 is realized lower if the character is preceded by a tone 3 character than preceded by any other tone. For a tone 4 character, there is a deep valley f0 contour shape when following a tone 3. In my experiments, the tone

[1]What people call "tone 0" or "neutral tone" is parasitic and cannot be treated equally as the 4 main lexical tones.

patterns of the last two characters are 1-4, 2-4, 3-4 and 4-4, which result in a noticeable lower tone 4 in the third sentence, and some listeners use that cue to identify the preceding target character as tone 3, as is shown in table2.

All of the complicated factors about tone 3 leading to the fact that the significant area of tone 3 is vague and highly various across participants. The untypical significant area pattern is the main cause of the unsatisfying performance of the classifier.

5 Conclusion and Future Work

Using the speech snippets that are the combination of a Mandarin sentence and random "bubble noise" and are labeled with the tone of the target character perceived by native Mandarin listeners, I build a convolutional neural network model to predict if a single snippet can be correctly identified. Three models except for tone 3 beat the baseline with satisfying performance. The prediction of tone 3 is harder as expected, because tone 3 is involved with many complicated factors, and even human predictions are not as stable as the other three tones.

This project is part of the pilot study for my thesis project, and there are more to explore. Tone perception, especially in noisy environment, is comparatively more difficult for L2 Mandarin learners than Native speakers. It is meaningful to compare how the two groups perceive tones differently, and how machines predict differently. Within all the wrongly identified speech, some are identified with a wrong tone, while others are absolutely unintelligible. There also tends to be some patterns of the errors. One observation is that the most common error for tone 2 is tone 3, while the most common error for tone 3 is tone 2, the second common error for tone 3 is tone 4. The reasons for error patterns should be explained with future research.

References

Yuen Ren Chao. 1968. *A Grammar of Spoken Chinese*. University of California Press.

Jiyoung Choi. 2018. Speech perception in "bubble" noise: Korean fricatives and affricates by native and non-native korean listeners. Master's thesis, City University of New York, Graduate Center, 365 Fifth Avenue, New York, NY 10016.

Joost M. Festen and Reinier Plomp. 1990. Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America*, 98:1725–736.

Jingyuan Huang and Lori L. Holt. 2009. General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, 125:3983.

Siyun Liu and Arthur G. Samuel. 2004. Perception of mandarin lexical tones when f0 information is neutralized. *Language and Speech*, 47(2):109–138. PMID: 15581188.

Michael I Mandel, Sarah E Yoho, and Eric W Healy. 2016. Measuring time-frequency importance functions of speech with bubble noise. *Journal of the Acoustical Society of America*, 140:2542–2553.

Odette Scharenborg. 2007. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49.

Xiaonan Susan Shen and Maocan Lin. 1991. A perceptual study of mandarin tones 2 and 3. *Language and Speech*, 34(2):145–156.

Yi Xu. 1997. Contextual tonal variations in mandarin. *Journal of Phonetics*, 25.

Bei Yang. 2010. *A Model of Mandarin Tone Categories – A Study of Perception and Production*. Ph.D. thesis, University of Iowa.