Automatically Detecting Reduced-formed English Pronunciations by Using Deep Learning

Lei Chen[1]**, Chenglin Jiang, Yiwei Gu, Yang Liu**[2]**, Jiahong Yuan**[3]
[1]Rakuten Institute of Technology (RIT)
[2]Alexa AI, Amazon
[3]Baidu Research
LAIX Inc. Shanghai China

Abstract

Reduced form pronunciations are widely used by native English speakers, especially in casual conversations. Second language (L2) learners have difficulty in processing reduced form pronunciations in listening comprehension and face challenges in production too. Meanwhile, training applications dedicated to reduced forms are still few. To solve this issue, we report on our first effort of using deep learning to evaluate L2 learners' reduced form pronunciations. Compared with a baseline solution that uses an ASR to determine regular or reduced-formed pronunciations, a classifier that learns representative features via a convolution neural network (CNN) on low-level acoustic features, yields higher detection performance. F-1 metric has been increased from 0.690 to 0.757 on the reduction task. Furthermore, adding word entities to compute attention weights to better adjust the features learned by the CNN model helps increasing F-1 to 0.763.

1 Introduction

The term "reduced forms" refers to the phenomenon of phonological simplification and variation commonly observed in connected speech of native speakers (Brown and Kondo-Brown, 2006; Khaghaninezhad and Jafarzadeh, 2014; Cangemi et al., 2018). In phonetic reduction, "segments may be shorter, less clearly articulated, or absent compared to canonical or dictionary forms" (Cangemi et al., 2018). Reduced-formed pronunciations appear in daily English communication among native speakers (Johnson, 2004).

On one hand, the existence of reduced form challenges second language (L2) learners in their listening comprehension (Norris, 1995). On the other hand, L2 learners often face great challenges on producing reduced forms so that they can sound more close to native speakers. English as a second language (ESL) teachers have realized the importance of specially training L2 learners on understanding reduced form pronunciations to improve their listening comprehension skills. (Yeh et al., 2017) is such an effort of designing an app to train students accordingly. Compared to the emphasis on reduced forms when training the listening comprehension skills, the effort on training the specific skills on the production of reduced forms is still limited. Most computer aided language learning (CALL) tools focus on training regular form pronunciations and do not provide adequate supports to L2 learners' demands on the production of reduced forms.

Hence, in this paper, we will report on our initial effort of using a deep learning based classification method to detect L2 learners' reduced form productions. Note that the detection is the first required step for creating a training tool that can generate feedback and provide interventions to cultivate L2 learners' specific skills.

2 Previous research

Reduced forms have been actively investigated in phonetics. For example, (Ernestus and Warner, 2011) introduced reduced-formed pronunciation variant phenomenal in phonetics. It pointed out that such variations are quite common in different languages in their casual conversation conditions. (Jurafsky et al., 1998) investigated English function words' reduced forms in the Switchboard corpus and found that a high percentage of reduced forms appears in the telephone conversations. Also, the authors investigated possible reasons causing reduced forms, such as words' frequencies.

(Wong et al., 2017) examined the role of the perception of reduced forms (e.g., contraction, elision, assimilation) of English words in connected speech comprehension and the phonological skills underpinning reduced forms perception. This study

[0] The work was conducted while Lei, Jiahong, and Yang were working in LAIX Silicon Valley AI Lab

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 22 - 26
July 15, 2022 ©2022 Association for Computational Linguistics

delivers a clear message to ESL teachers that the ability of perceiving reduced pronunciation variants is important for L2 listening comprehension skills. There are some emerging technical works on helping L2 learners' perception of reduced forms. For example, (Yeh et al., 2017) reports on an Android App to use authentic native connected speeches as material to teach.

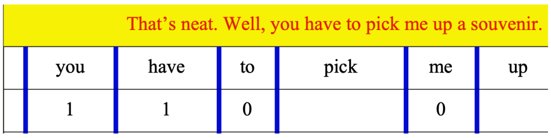
Reduced forms sometimes are produced as variants to formal forms. Based on this fact, methods that can distinguish pronunciation variants in ASR can be used to detect the existence of reduced forms. (Strik and Cucchiarini, 1999) systematically surveyed the methods for recognizing pronunciation variants. In a widely used approach, extended recognition network (ERN) (Qian et al., 2016; Harrison et al., 2009), extra decoding paths are added to represent pronunciation variants on top of the regular paths built on formal forms' pronunciations. However, reduced forms can sometimes occur without pronunciation variants. Hence, solutions working on broader cases are worth investigating.

3 Data

In this paper, we focused on the two types of reduced forms, i.e., *reduction* and *liaison*. The first refers to changing pronunciation from its formal form on individual words while the second term refers to co-articulation among adjacent words.

In 2019, Company-X released a new product for training various specific pronunciation skills, including reduction and liaison, in its main English learning mobile App. This product has already been used by a large number of Chinese English learners. From the audio samples collected in this product, we built up our own research data set. We sampled speech files from a large group of English learners from different locations in China. When sampling L2 learners' spoken responses, we used pronunciation scores automatically rated by Company-X AI-based pronunciation scoring system to include learners from diverse levels.

On $8,570$ practice audio samples for the reduction skill, $seven$ human raters annotated whether learners produce correct reductions or not on the required words. During rating, the annotators considered three aspects, including energy, rhythm (duration and its connection to context words), and pronunciation variations. These raters are high-level non-native English speakers and doing linguistics and phonetics annotation as their full-time jobs. For each audio sample, if at least four raters agree on one label, this label will be used to be the sample's final decision. Otherwise, the sample will be treated to be too challenging for human annotators and will be excluded from the experiments. The entire rating was done in two stages by using two groups of double raters. For each stage, a kappa set was used to measure two raters' rating consistence. In the first stage, the rating agreement was $\kappa = 0.63$. Then, raters obtained more training on understanding the rating guideline before going to the second stage of rating. The agreement measurement on the second stage has been increased to $\kappa = 0.79$. Figure 1 shows the annotation interface in Praat software. We can see that for selected words, e.g., "you", "have", "to", and "me", human raters used annotation tiers to label their decisions. "1" denotes reduction while "0" denotes formal form.



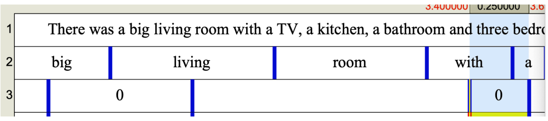
That's neat. Well, you have to pick me up a souvenir.

you | have | to | pick | me | up

1 | 1 | 0 | | 0 |

Figure 1: Annotation interface for reductions in Praat

On $4,027$ practice audio samples for the liaison skill, three human annotators marked whether some word pairs are spoken as liaisons or not. Figure 2 shows the annotation interface in Praat. For the word pairs "big living" and "will a", annotators mark 1 indicating two words are spoken in a connected way or use 0 indicating two words are spoken in their formal forms. For each audio sample, the majority voting results among the three annotators were used as final labels. Among the three raters, their between-rater agreement values are 0.74, 0.81, and 0.82 respectively. This shows that judging liaison is relatively easy compared to judging reductions. Table 1 summaries the label counts of the two data sets.



There was a big living room with a TV, a kitchen, a bathroom and three bedro

big | living | room | with | a

0 | 0

Figure 2: Annotation interface for liaison detection in Praat

Reduced form type	#Yes	#No	#Total
reduction	$3,610$	$7,953$	$11,563$
liaison	$1,334$	$2,693$	$4,027$

Table 1: Statistics of the two reduced form data sets

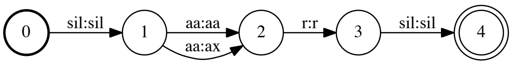


Figure 3: Decoding network for the word “are" with formal form /ar/ and reduced form /r/

4 Models

4.1 ASR based method

For words (including single words or multi-word with reduced forms, e.g. “got you" to “gotcha") with pronunciation variations, a typical solution is using an ASR system with two distinct pronunciation entries in its dictionary to determine real pronunciation on the fly. For example, “are” can be pronounced as its formal form /ar/ or its reduced form /r/. Here, we first use a forced alignment step to locate words being considered. Then, a recognition network with paths mapping to two pronunciation forms is used to decode the audio portion to find the more possible path. Figure 3 shows one concrete example of a decoding network for the word “are".

The automatic speech recognition (ASR) system was built with the Kaldi open-source toolkit. This model is a 9-layer time delayed neural network (TDNN) using Mel frequency cepstrum coefficient (MFCC) acoustic features from the current frame plus the previous and following 5 context frames. The ASR model was trained on our in-house read-aloud corpus containing about $2,500$ hours of native and non-native speech files. The ASR system achieved a word error rate (WER) of 9% on learners' speech.

4.2 CNN based method

ASR based method was based on an assumption that pronunciation variations always occur in the reduced forms. However, some reduced forms may only show in low energy levels and shorter duration. To address this limitation, we investigated building a classifier to predict a audio segment's pronunciation form directly,

Pronunciation in a reduced form is a complicated process. To obtain effective representations, we conducted an automatic feature learning by utilizing a convolution neural network (CNN) model (Abdel-Hamid et al., 2014). For the reduction detection task, we used MFCC feature sequence over each audio segment being considered. For the liaison detection task, we used the MFCC feature sequence starting from the last phoneme of the starting word to the first phoneme of the ending word in each adjacent word pair being considered. Also, these two phonemes were connected to form a token to be the “word entity" associated with this word pair. In each reduced form detection task, all of the audio portions were padded to the same length. For example, for the reduction detection task, all portions were padded to 0.5 second long.

The **librosa** (McFee et al., 2015) V0.7 audio signal processing Python package was used to extract MFCC ($n = 40$) features. For example, in the reduction detection task, each input feature (on a word) takes a shape of 16×40. Then, we sent these tensors to two CNN blocks, each block contains a 1D CNN (filters numbers are 128 and 256 respectively) and a batch normalization (BN) (Ioffe and Szegedy, 2015) layer. From the second CNN block's output, global max pooling and Dropout (Srivastava et al., 2014) layers were used to convert learned features to be vectors with a dimension of 256. At last, the learned features went through a fully connected (FC) layer using a sigmoid activation function to obtain reduced form prediction binary output.

All of the audio clips used in this study were collected from learners when they practiced on a set of pre-defined words. We noticed that learners' reduction production varied among these word entities. Therefore the word entity's prior information is expected to be useful for modeling learners' production behaviors. To incorporate word entity cues in the reduced form prediction, we utilized word entities one-hot representations to compute feature attention weights so that for each specific word entity, a different feature weighting plan can be learned in our CNN model. The learned feature from CNNs is denoted as $\mathbf{F} = \{\mathbf{f_t}\}$ where $0 <= t <= 255$. $\mathbf{V_i}$ is the one-shot encoding vector for the word w_i among all $|V|$ pre-defined words for testing L2 speakers' reduction production capabilities. We use a linear mapping $\mathbf{W}$, which is learned during model training, and a softmax activation to compute attention weights α_t. Then, an adjusted feature vector $\mathbf{S}$ is obtained by applying the attention weights on $\{\mathbf{f_i}\}$ element-wise.

$$A_i = \{a_t\} = W \times V_i \tag{1}$$

$$\alpha_{\mathbf{t}} = \frac{\mathbf{exp(a_t)}}{\sum_{\mathbf{k=0}}^{\mathbf{255}} \mathbf{exp(a_k)}} \tag{2}$$

$$\mathbf{S} = \{\alpha_{\mathbf{t}}\mathbf{f_t}\} \tag{3}$$

Figure 4 depicts our CNN models in details. Note that the left panel shows the model only considering audio information while the right panel shows how word entities were used to compute attention weights to adjust the learned features dynamically. The model was implemented by using Keras package (Chollet et al., 2015) on TensorFlow (Abadi et al., 2015).

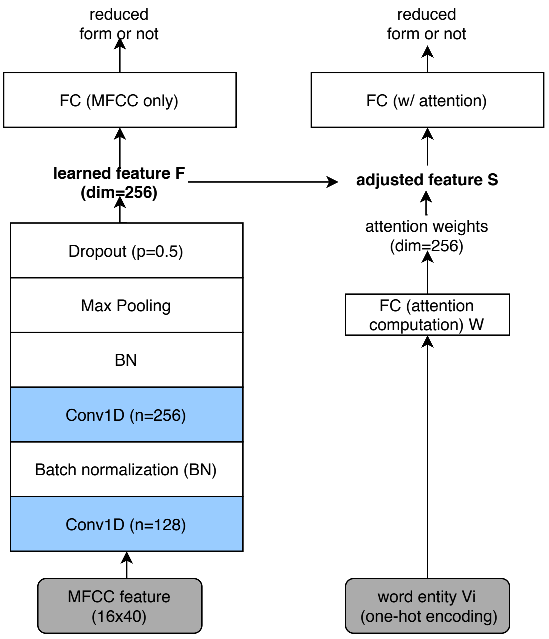


Figure 4: Using 1D CNN to learn features from MFCC for predicting whether a reduced form pronunciation occurs or not. The learned feature can be directly used (as in the left side) or adjusted by using attention weights based on word entities (as shown in the right side)

5 Experiments

Using the entire data set, we run our ASR based method to obtain reduced form predictions. When evaluating our classification based methods, we run cross-validation experiments. To compensate non-deterministic effects of using neural networks, we repeated our CV experiments 5 times and reported average performance among.

Regarding evaluating methods, we used standard metrics when evaluating binary classification, i.e., accuracy and F-1 score weighted by label percentage. The higher measurement metrics, the better the methods.

Method	Accuracy	F-1 (mean)
ASR	70.38%	0.690
CNN	76.19%	0.757
CNN + ATTN	77.01%	0.763

Table 2: Reduction detection results measured in both accuracy and F-1 score on different methods

Method	Accuracy	F-1 (mean)
ASR	70.85%	0.692
CNN	71.49%	0.719
CNN + ATTN	72.11%	0.720

Table 3: liaison detection performance measured in both accuracy and F-1 score on different methods

Table 2 reports on the experiment for the reduction detection task. CNN model shows improvements over the baseline ASR model, suggesting that CNN can automatically learn more indicative features from audio signals. When using attentions based on word entities to adjust the learned features, we can find further performance improvements (F-1 from 0.757 to 0.763).

Table 3 reports on the experiment result for the liaison detection task. Similar to what we found on the reduction task, the two methods using a CNN model to learn features automatically show improved performance than the method based on ASR decoding. Also, using attention weights computed based on phoneme-pairs is helpful.

6 Discussion

Reduced forms are commonly used by native speakers in their casual conversations. Because L2 learners mostly face formal forms in their language learning, perception and production of reduced forms in fact greatly challenges learners' listening comprehension and speaking capabilities. With a goal of building a training application on producing reduced-formed pronunciations, we conducted a research on automatically detecting reduced forms with a high accuracy. Following on the work of recognition of pronunciation variants, we firstly utilized an ASR decoding method to distinguish formal vs. reduced forms. To cope with reduced forms without obvious pronunciation variations, we then explored using a CNN model to learn distinguishable features from learner speech directly.

Our experiment results show that CNN method has improved performance over the ASR decoding method. Using word entities in the CNN model to compute attention weights to adjust the learned features is proven to be useful. Overall, on the two reduced form types, i.e., reduction and liaison, our CNN model has F-1 measurement about 0.763 and 0.720 respectively.

We envision that there are several research directions in future. Firs, so far, we only used CNNs to encode MFCC feature sequence, it is worthwhile trying some new encoding method, like Transformer. Second, human annotation on reduction still have a room to improve. We are hoping to continue increasing rating agreement to provide a even more solid research base. At last, a training module has been added into LAIX Liulishuo App. Based on real user data, it is worthwhile tracking whether provided training helps on learners' mastery of reduced forms.

References

Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, and Matthieu Devin. 2015. Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(10):1533–1545.

J. D. Brown and K. Kondo-Brown. 2006. Testing reduced forms. In *Perspectives on teaching connected speech to second language speakers*, pages 247–264.

Francesco Cangemi, Meghan Clayards, Oliver Niebuhr, Barbara Schuppler, and Margaret Zellers. 2018. Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation.

François Chollet et al. 2015. Keras. `https://keras.io`.

Mirjam Ernestus and Natasha Warner. 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39(SI):253–260.

Alissa M. Harrison, Wai-Kit Lo, Xiao-jun Qian, and Helen Meng. 2009. Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *International Workshop on Speech and Language Technology in Education*.

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.

Keith Johnson. 2004. Massive reduction in conversational american english. *Yoneyama, K and Maekawa, K. (Eds), Spontaneous Speech: Data And Analysis*.

Daniel Jurafsky, Alan Bell, Eric Fosler-Lussier, Cynthia Girand, and William Raymond. 1998. Reduction of English function words in Switchboard. In *Fifth International Conference on Spoken Language Processing*.

Mohammad Saber Khaghaninezhad and Ghasem Jafarzadeh. 2014. Investigating the Effect of Reduced Forms Instruction on EFL Learners' Listening and Speaking Abilities. *English Language Teaching*, 7(1):159–171.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.

Robert W. Norris. 1995. Teaching reduced forms: Putting the horse before the cart. In *English Teaching Forum*, volume 33, pages 47–50.

Xiaojun Qian, Helen Meng, and Frank Soong. 2016. A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(6):1020–1028.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Helmer Strik and Catia Cucchiarini. 1999. Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29(2-4):225–246.

Simpson WL Wong, Peggy PK Mok, Kevin Kien-Hoa Chung, Vina WH Leung, Dorothy VM Bishop, and Bonnie Wing-Yin Chow. 2017. Perception of native English reduced forms in Chinese learners: Its role in listening comprehension and its phonological correlates. *TESOL Quarterly*, 51(1):7–31.

Hsin-Yu Yeh, Yu-Tzu Tsai, and Chih-Kai Chang. 2017. Android app development for teaching reduced forms of EFL listening comprehension to decrease cognitive load. In *2017 International Conference of Educational Innovation through Technology (EITT)*, pages 316–321. IEEE.