Using deep learning to classify English native pronunciation level from acoustic information

Aozora Kobayashi[1,*] and *Ian* Wilson[1,**]

[1]CLR Phonetics Lab, University of Aizu, Tsuruga, Ikki-machi, Aizuwakamatsu, Fukushima-ken, Japan

Abstract. The main purpose of this research is to test the use of deep learning for automatically classifying an English learner's pronunciation proficiency, a step in the construction of a system that supports second language learners. Our deep learning dataset consists of 28 speakers – ranging in proficiency from native to beginner non-native – reading the same 216-word English story. In the supervised deep learning training process, we first label the English proficiency level of the data, but this is a complicated task because there are a number of different ways to determine someone's speech proficiency. In this research, we focus on three elements: foreign accent, speech fluency (as measured by total number of pauses, total length of pauses, and speed of speech) and pronunciation (as measured by speech intelligibility). We use Long Short-Term Memory (LSTM) layers for deep learning, train a computer on differently labeled data, test a computer on separate data, and present the results. Features used from audio data are calculated by Mel-Frequency Cepstrum Coefficients (MFCCs) and pitch. We try several combinations of parameters for deep learning to find out what settings are best for our database. We also try changing the labeling method, changing the length of each audio sample, and changing the method of cross-validation. As a result, we conclude that labeling by speech fluency instead of by speech intelligibility tends to get better deep learning test accuracy.

1 Introduction

1.1 Motivation and goal

English is an important tool for communication in an increasingly global society, and many non-native speakers study English-as-a-Foreign-Language (EFL). However, it is sometimes difficult for second language (L2) speakers to perceive and produce pronunciation differences between other speakers (including natives) and themselves. Although this is not necessarily a problem for communication, such differences often do result in communication breakdown.

Japanese English learners have common tendencies in their English pronunciation errors [1]. For instance, Japanese learners of English tend to mistake the pronunciation of the high back lax vowel /ʊ/ because it is not an underlying phoneme in Japanese words. So, a minimal pair such as "fool" and "full" may sound the same to them. Moreover, in general, L2 speakers' speed of speech is slower than native speakers, because they are overly careful when speaking in English. In this way, there are many English pronunciation differences between native speakers and non-native speakers, some of which lead to communication problems.

Based on this situation, the ultimate goal of our research is constructing a system that supports pronunciation study for EFL learners. A system that can give feedback about differences between L2 speakers' and native speakers' English pronunciation would be a valuable pedagogical tool. Aside from pedagogical purposes, such a system might be useful for forensic phonetics and security, when people need to make positive identifications of speakers. To construct such a system, it is helpful to automatically detect errors in speech and judge the level of English pronunciation (see [2] for a good overview).

However, categorizing a speaker's pronunciation proficiency is not a straightforward task. To categorize speakers, the problem is determining what factors cause people to think "This speech sounds like a native speaker" or "Person A sounds more proficient than person B". Ratings of language proficiency can be measured in many different ways [3], [4], [5].

As a step towards making a pedagogical system as described above, we first use deep learning for categorizing non-native speakers' English speaking proficiency. Deep learning is a type of machine learning that uses deep neural networks (DNNs). One of the merits of using deep learning is that a machine can choose the salient features from many features in the data. To investigate speaking proficiency of L2 speakers, this method seems more appropriate than general machine learning because speech has so many features, and proficiency has so many indicators. Therefore, we chose deep learning for the main method in this research.

Moreover, to get a large enough training dataset for use with deep learning, we used (and added to) an existing database in our lab. There are 4 types of data in our

*e-mail: m5221101@u-aizu.ac.jp
**e-mail: wilson@u-aizu.ac.jp

database: audio data, ultrasound tongue video data, side-view face video data, and front-view face video data [6]. In this paper, we focus on only the audio data.

1.2 Deep learning

As mentioned above, deep learning is a type of machine learning that uses DNNs. A deep neural network (DNN) is an artificial neural network with multiple layers between the input and output layers. The type of neural network that we choose is important. For example, Convolutional Neural Network (CNN) is suitable for image data, and Recurrent Neural Network (RNN) is better than CNN for data with a time series [7].

In our research, Long Short-Term Memory (LSTM) is most frequent. LSTM is one type of RNN architecture. A common LSTM unit is composed of a *cell*, an *input gate*, an *output gate*, and a *forget gate*. The cell remembers values over arbitrary time intervals and the gates control the information flow into and out of the cell. Thus, an LSTM network is very suitable for classification and prediction with time-series data.

In machine learning, the method the computer uses to learn during training is important, specifically unsupervised learning or supervised learning. Unsupervised learning means using a training dataset without pre-existing labels – the computer determines categories on its own – whereas supervised learning means using a training dataset with pre-existing user-defined labels [7].

In the case of labeling for training deep learning models, one factor that could affect results is label noise – the problem of not being able to clearly label some of the data or the problem of a lack of agreement between experts on a given label [8]. If one trains DNNs on noisy labeled data, overfitting can occur, which lowers the effectiveness of the deep learning [9]. In the past, some researchers tried to make DNNs that overcome strong label noise [10] and estimate probabilities of label noise.

In past research, many researchers have tried to automatically recognize pronunciation differences between native speakers and non-native speakers by using a variety of ways (for an overview, see [2]). One example is using tongue trajectories of ultrasound tongue images [11]. In that research, a cosine convolution filter was used to extract features from ultrasound tongue images. Native and non-native speakers' tongue motion was compared and pronunciation errors could be automatically detected, especially for /k/, /ʃ/, and /r/.

Other deep learning research has used not only audio speech data, but also ultrasound image data [12] and the combination of the two. In that research, the authors compared (1) the accuracy of deep RNN by using only audio data to (2) the accuracy of using both audio and point-tracked tongue information. They found that using features of both types of data can get higher accuracy for deep learning than using features of only one data type.

2 Database for deep learning

2.1 Participants

We used data from a total of 28 participants in this study – 23 who were from a preexisting database [6] and an additional 5 participants whose data we collected in exactly the same way as in the past. The former 23 were used in Trials 1 & 2, and the entire 28 were used in Trials 3 & 4. Table 1 shows information about the participants: their sound (WAV) file identification number(s) for internal use, age, gender, and English proficiency (TOEIC) score, if available. The 5 participants who were added after Trials 1 & 2 have an asterisk after the participant number.

Table 1: Participant list (ordered by TOEIC score)

Participant #	WAV file #	Age	Gender	TOEIC score
1	618	32	M	990[1]
2	621	33	F	965
3*	635, 636	28	M	960[2]
4	619	30	F	920
5	617	—	F	910
6	622	60	F	900[3]
7	616	49	F	900[3]
8*	632, 633	44	M	900[3]
9*	639, 640	21	F	810
10*	637, 638	22	M	765
11	620	32	F	685
12	598	23	M	650
13	597	24	M	600
14*	641, 642	19	M	595
15	603	20	M	560
16	605	24	M	510
17	599	23	M	450
18	612	24	M	440
19	613	20	M	425
20	604	24	M	400
21	606	24	M	400
22	610	19	M	400
23	600	19	M	395
24	607	19	M	375
25	609	18	F	375
26	602	23	M	350
27	615	23	M	350
28	608	21	M	225

* participant added after Trials 1 & 2
[1] actually native speaker (USA)
[2] converted from TOEFL iBT=98
[3] converted from Eiken Level 1

2.2 Data collection

2.2.1 Apparatus

The apparatus used to collect data can be seen in Figure 1 and is described in detail in [6]. Participants sat on a chair in front of a laptop that displayed stimuli using Powerpoint

slides. Ultrasound tongue image video was mixed with audio and exported to an iMac computer. Participants needed to wear a helmet holding the probe of the ultrasound machine to prevent the probe from shifting relative to the skull during speech. We recorded participants' speech at a 48,000 Hz sampling rate, using a DPA 4080 cardioid lapel microphone and a Korg MR-1000 recorder. Speakers repeated the same stimuli passage twice, because we recorded ultrasound movies in two separate planes: midsagittal and coronal.

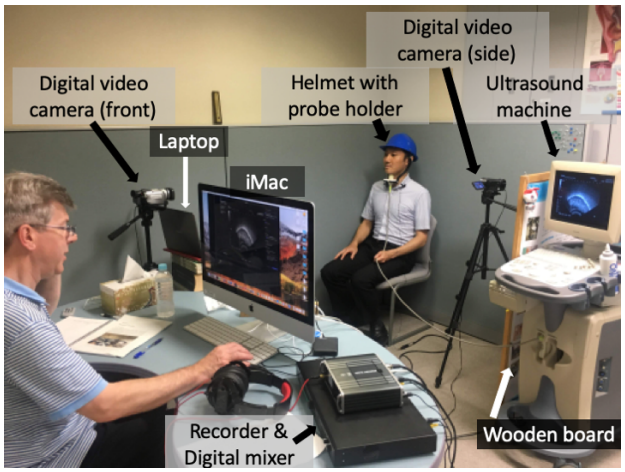

Figure 1: Data collection apparatus in the CLR Phonetics Lab

2.2.2 *Stimuli*

As English stimuli, we used a version of the famous Aesop fable "The Boy Who Cried Wolf" [13]. The sentences below show the whole 216-word Wolf Passage. Note that this stimuli has at least three clear instances of each of the English monophthong vowels, most of the diphthongs and consonants, and even some minimal pair words such as "fist"/"feast" and "raising"/"racing". There are 134 unique words in the passage, in a great range of phonetic environments.

> There was once a poor shepherd boy who used to watch his flocks in the fields next to a dark forest near the foot of a mountain. One hot afternoon, he thought up a good plan to get some company for himself and also have a little fun. Raising his fist in the air, he ran down to the village shouting "Wolf, Wolf". As soon as they heard him, the villagers all rushed from their homes, full of concern for his safety, and two of his cousins even stayed with him for a short while. This gave the boy so much pleasure that a few days later he tried exactly the same trick again, and once more he was successful. However, not long after, a wolf that had just escaped from the zoo was looking for a change from its usual diet of chicken and duck. So, overcoming its fear of being shot, it actually did come out from the forest and began to threaten the sheep. Racing down to the village, the boy of course cried out even louder than before. Unfortunately, as all the villagers were convinced that he was trying to fool them a third time, they told him, "Go away and don't bother us again". And so the wolf had a feast.

2.3 Data cleaning

When training a computer on a dataset for deep learning, it is important to have as clear a signal as possible – whether it is audio, visual, or other. Since participants sometimes made whole-word mistakes in their pronunciation of the stimuli, and since the signal intensity was not consistent across participants, we had to clean the data before deep learning. We used Praat software [14] to carry out the following steps for data cleaning: (1) extracted the required part of the English stimuli, (2) scaled intensity to a new average of 70 dB, (3) converted audio from stereo to mono, and (4) downsampled audio sampling rate from 48,000 Hz to 24,000 Hz. In step 2, the root-mean-square amplitude of the sound will come to lie 70 dB above the assumed auditory threshold of 0.00002 Pa.

2.4 Data labeling

As mentioned in Section 1.2, before using the supervised learning type of deep learning, we needed to label the data. The problem is that there are so many ways to categorize participants into English native speech levels: foreign accent, pronunciation intelligibility, fluency, grammar, vocabulary, pragmatics, etc. To measure fluency, past researchers have used speech rate in words per minute (WPM) [15] and number and length of pauses [16]. Measurement of speech intelligibility is often done by having listeners transcribe spoken speech and seeing how many words they get correct [17].

When determining the final labels in the supervised deep learning training process, we decided to try defining labels using different factors such as perceived foreign accent, fluency, and intelligibility. These different labeling methods will be described in more detail in Section 3.

3 Deep learning data analysis and results

3.1 Trial 1: Two-second audio segments

In this trial, we used audio data from 23 participants. Before extracting features from the data, we separated the audio into many 2-second files because the length of the original sound files were too long to process in computer memory. We allocated 1139 files (about 80%) for training and 310 files (about 20%) for testing. As for features, we used 13-dimensional Mel-Frequency Cepstrum Coefficients (MFCCs) and pitch (fundamental frequency (f0)). MFCCs are coefficients that collectively make up the Mel-frequency cepstrum. To calculate both MFCCs and f0, we

used the MATLAB function "HelperComputePitchAndMFCC" [18]. After calculating MFCCs and f0 features, we normalized the features to be from -1 to 1. Adding the pitch feature to the MFCCs, the resultant input data was 14-dimensional vectors.

We categorized the 23 participants' data into groups, labeling based on judgements of degree of nativeness obtained from 8 native or near-native EFL professors at the University of Aizu. At first, we made a questionnaire using a Google form, a free tool provided by Google. Professors listened to each person's first two sentences from the Wolf Passage. The order of the files was randomized. After listening to an audio sound file from a given speaker, they rated the speaker's English pronunciation level on a scale from 1 to 7, where 7 meant that it sounds like an English native speaker and 1 meant that it sounds like a very poor non-native English speaker. After collecting all results, we calculated an average level for each speaker and divided the 23 participants (speakers) into 5 levels.

After that, the machine trained on the labeled data and output the test results. Settings for this trial and all the others are shown in Table 2. Optimizer is the algorithm for operating parameters from input layer to output layer. Initial learning rate is the step size in optimization for weight adjusting. The size of MiniBatch is the number for choosing training data randomly. If we set MiniBatch size to be 25, training data is split into 25 sets randomly. Epochs means the number of times that the machine trained on the whole training dataset. The purpose of L2 regularization is to prevent overfitting, which is a failure to fit additional data to the training data or to predict future observations reliably. The activation function is a function that is included in the optimization step.

After training the computer on 80% of the sound files, we used the remaining 20% to test how well the computer could learn to classify the sound files into proficiency levels. The test results for Trial 1 are shown in Figure 2. The row labels in this figure are the actual given labels of the test data, and the column labels are ones calculated by the neural network. So for example, in the top-left cell of the table, 6159 samples of Level 1 (labeled as the lowest proficiency from the professors' judgments) were accurately classified as Level 1 by the computer. In the top-right cell of the table, 848 samples of Level 1 were inaccurately classified as Level 5 (the highest proficiency) by the DNN. In the line above the 25 cells, accuracy can be seen to be 25.246% – only about 5% better than random chance, which would be 20%. In an attempt to obtain a higher accuracy, we decided to use different lengths of audio segments in Trial 2.

3.2 Trial 2: One-second and 0.1-second audio segments

In the previous deep learning trial, audio files were divided into many 2-second samples for deep learning. However, past researchers have used other lengths of samples [12]. Therefore, in Trial 2, we used three different sample lengths for the audio files, to discover what sample length could achieve the best test accuracy for our data.

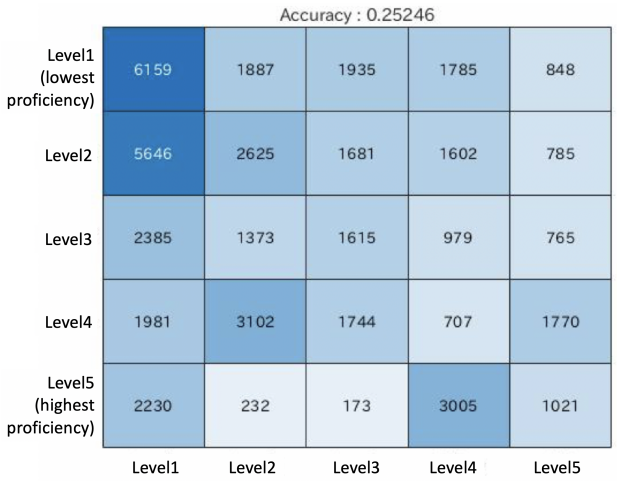

Figure 2: Test results and accuracy in Trial 1. Row labels are the actual given labels of the test data, and column labels are ones calculated by the DNN

The three sample lengths were 100 milliseconds, 1 second, and 2 seconds. Table 3 shows the number of files used in training and testing for each sample length. In Trial 2, the labeling method, the features that were used and the amount of data were the same as in Trial 1. Because of the increase in the number of sound files, we had to decrease the setting for maximum number of epochs, otherwise the training portion would have taken much too long (i.e., many days of processing).

The results showing the highest test accuracy for each of the three sample lengths are shown in Figure 3. As can be seen, the test accuracy of 100-millisecond sample lengths (25.9%) was slightly higher than the other two sample lengths. Therefore, in Trials 3 and 4, we used 100-millisecond sample audio data for deep learning.

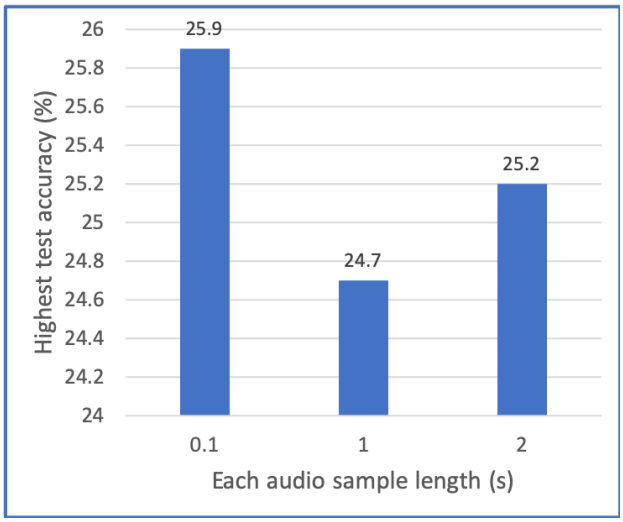

Figure 3: Highest test accuracy obtained for each sample time length in Trial 2

Table 2: Participants, stimuli segments, labeling, and deep learning parameters used in Trials 1–4

Parameter	Trial 1	Trial 2	Trials 3 & 4
# of participants	23	23	28
Audio segment length (s)	2	2, 1, 0.1	0.1
Labeling	Nativeness	Nativeness	Fluency & intelligibility
Optimizer	Momentum (SGD)	Momentum (SGD)	Momentum (SGD)
Initial learning rate	0.01	0.01	0.1
MiniBatch size	25	25	64
Max epochs	70	30	30
L2 regularization	0.00001	0.00001	0.0001
Classification output activation	Softmax	Softmax	Softmax

Table 3: The number of training and testing files for each sample length in Trial 2

Sample length (s)	Total # of training wav files	Total # of testing wav files
0.1	22922	6249
1.0	2279	621
2.0	1139	310

3.3 Trial 3: More participants and a variety of labeling methods

In Trials 1 and 2, two problems existed in our database: (1) not enough high-proficiency speakers meant that our database was unbalanced, and (2) subjective judgments caused excessive noise in the labeling of data. For these reasons, after adding five new higher-proficiency speakers' audio data, in Trial 3 we compared results of different labeling methods (the labeling used for categorizing English pronunciation proficiency level). The labeling methods in Trials 3 and 4 are as follows: (1) number of pauses in the whole speech, (2) total length of pauses in the whole speech, (3) speed of speech as measured by WPM, and (4) labeling of data by intelligibility calculations obtained from a native speaker transcription task in Amazon Mechanical Turk (AMT).

3.3.1 Labeling data by acoustic fluency information (Trials 3 & 4)

Fluency of speech is one factor in judging someone's L2 pronunciation. In this research, we calculated and labeled fluency using three methods: (1) total number of pauses in the whole speech, (2) total length of pauses in the whole speech, and (3) speed of speech in WPM. In the first two methods, a pause was defined as ≥400 ms, following [19].

In this way, we divided the 28 participants into 4 levels. We changed the number of levels from 5 in the first two trials to 4 here because 4 is a divisor of 28, so each of our groups could have an equal number of participants. Unbalanced group numbers in Trials 1 and 2 were unavoidable, but could have negatively affected deep learning results. Table 4 shows proficiency group labeling for total number of pauses, total length of pauses, and speed of speech in WPM. Figure 4 shows the same data as Table 4, but plotted for each individual participant.

3.3.2 Labeling data by intelligibility scores from Amazon Mechanical Turk (Trials 3 & 4)

Another way of objectively labeling the data was using the AMT system to get many listeners' intelligibility scores. AMT is one of the services offered by Amazon Web Services. AMT is a social marketplace for tasks that require talented human intelligence. AMT is useful for L2 speech judgements by crowdsourcing, because we can easily decide the requirements of participants [20]. To evaluate intelligibility, we used a dictation test [21], which is used to determine how many words listeners are able to correctly identify in the speakers' speech. For judging this factor, we used AMT because of the potential for a greater number of available participants. We set a requirement that AMT participants be native English speakers resident in the USA.

We used HTML for making the survey page and based it on past research [22]. Overall, there were 56 randomly-chosen questions for each AMT worker – two sound files from each of the 28 participants – plus one practice question with a native English speaker's audio file. At first, we divided the wolf-passage stimuli into 56 short phrases. Table 5 shows all the phrases. Second, we made mp3 files of all 56 phrases from each participant in the database. After that, we pseudo-randomized the order such that the first half (28 questions) for a given worker were randomly selected from phrases 1 to 28, with each phrase being a different speaker. The second half (28 questions) for a given worker were randomly selected from phrases 29 to 56, again with each phrase being a different speaker. Doing this enabled us to make sure that AMT workers listened to each database participant twice in this survey and each AMT worker heard phrases and participants in a random order. A flowchart of how we sorted pseudo-randomly is shown in Figure 5. Workers could listen to each of the 56 audio files up to two times and had to type exactly what they heard each participant say.

After making the HTML file, we registered our questionnaire in AMT as a requester, and we set it to close in 10 days. As a qualification of our questionnaire, we required that the workers' (listeners') locations be within

Table 4: Proficiency group labeling, for Trials 3 & 4, for all participants by each of three measurements of fluency (total # of pauses, total length of pauses, speed in words/minute) and one measurement of pronunciation (% intelligibility calculated from 54 native listeners' transcriptions)

Proficiency Ranking (high to low)	Proficiency Group label	Partic. #	Total # of pauses	Partic. #	Total length of pauses (s)	Partic. #	Speed (WPM)	Partic. #	Word intelligibility from AMT (%)
1		7	12	1	23.2	10	156.9	8	93.5
2	↑	1	17	15	23.2	1	152.4	7	92.2
3		15	19	12	23.5	7	134.5	4	89.6
4	4	6	19	10	23.8	3	127.8	1	89.4
5		10	19	6	27.0	8	125.4	5	89.0
6	↓	4	20	5	28.5	11	125.3	12	85.7
7		8	20	7	28.5	15	124.3	15	85.5
8		12	21	11	29.5	6	115.2	6	84.3
9	↑	5	21	2	31.1	26	111.3	14	83.1
10		2	21	13	31.1	4	110.9	3	81.3
11	3	11	23	8	31.2	12	109.6	11	78.3
12		9	26	18	31.3	5	108.1	26	76.8
13	↓	13	28	26	33.9	16	107.3	27	76.3
14		26	28	3	35.3	13	105.1	10	72.2
15		3	29	22	35.3	23	103.7	17	72.1
16	↑	16	32	16	37.0	9	103.3	13	72.0
17		18	32	27	37.8	18	102.1	2	70.2
18	2	14	32	17	39.1	22	101.0	9	68.6
19		25	33	4	39.6	21	99.8	18	66.8
20	↓	27	33	25	40.5	17	99.2	24	64.4
21		23	34	9	45.0	27	95.1	16	64.1
22		22	35	24	45.8	2	94.2	22	64.1
23	↑	24	36	23	48.2	24	92.9	21	62.9
24		19	37	19	49.5	25	90.3	19	62.4
25	1	17	39	21	53.1	19	87.8	28	58.4
26		21	41	20	60.3	14	87.8	20	55.7
27	↓	20	46	14	62.5	20	84.9	23	54.5
28		28	79	28	96.9	28	64.0	25	36.4

the USA, because most English learning systems in Japan are based on American English, and the location is easy to set that way in AMT. We also set the qualification that workers must be master-level workers to raise the probability that we got only dedicated workers. Master level is an Amazon-defined level meaning that workers have received a very high percentage of positive feedback from requesters in the past. In 10 days, we could get answers from 55 workers. Looking at the answers, one worker's results were considered unreliable, because that worker completely mis-transcribed the native English speaker's clear pronunciation, which we had added as a practice part in the questionnaire. Therefore, we used the remaining 54 workers' answers to determine the group labeling. After checking answers, we paid 2 dollars to each worker, except for the worker who we judged to be unreliable.

Based on 54 workers' answers, we calculated the score of each participant speaker. We evaluated speakers by comparing the workers' typed results and the correct answer for each word. For measuring intelligibility, whole-word identification rather than phoneme identification is important, so we scored intelligibility based on the number of correct whole words. Then for each speaker, we calculated their AMT accuracy by dividing the number of workers' correct answer words by the total number of words in the phrases. After that, we divided the speakers into 4 equal groups based on the results, labeling them from level 1 (low proficiency) to level 4 (high proficiency). The final proficiency labels according to AMT workers' transcriptions are shown in Table 4, and Figure 4 shows the same data as Table 4, but plotted for each individual participant.

We then compared test accuracies for each labeling method, in order to find the best way to label for classifying English pronunciation levels. The numbers of files per label is shown in Table 6. The deep learning training network layers were: input > LSTM > dropout > fully connected > softmax > classification. The test accuracy results for each labeling method are shown in Figure 6. As can be seen, when data was labeled by elements of speech fluency (total number of pauses, total length of pauses and WPM) it got better test accuracy than when data was labeled by AMT intelligibility.

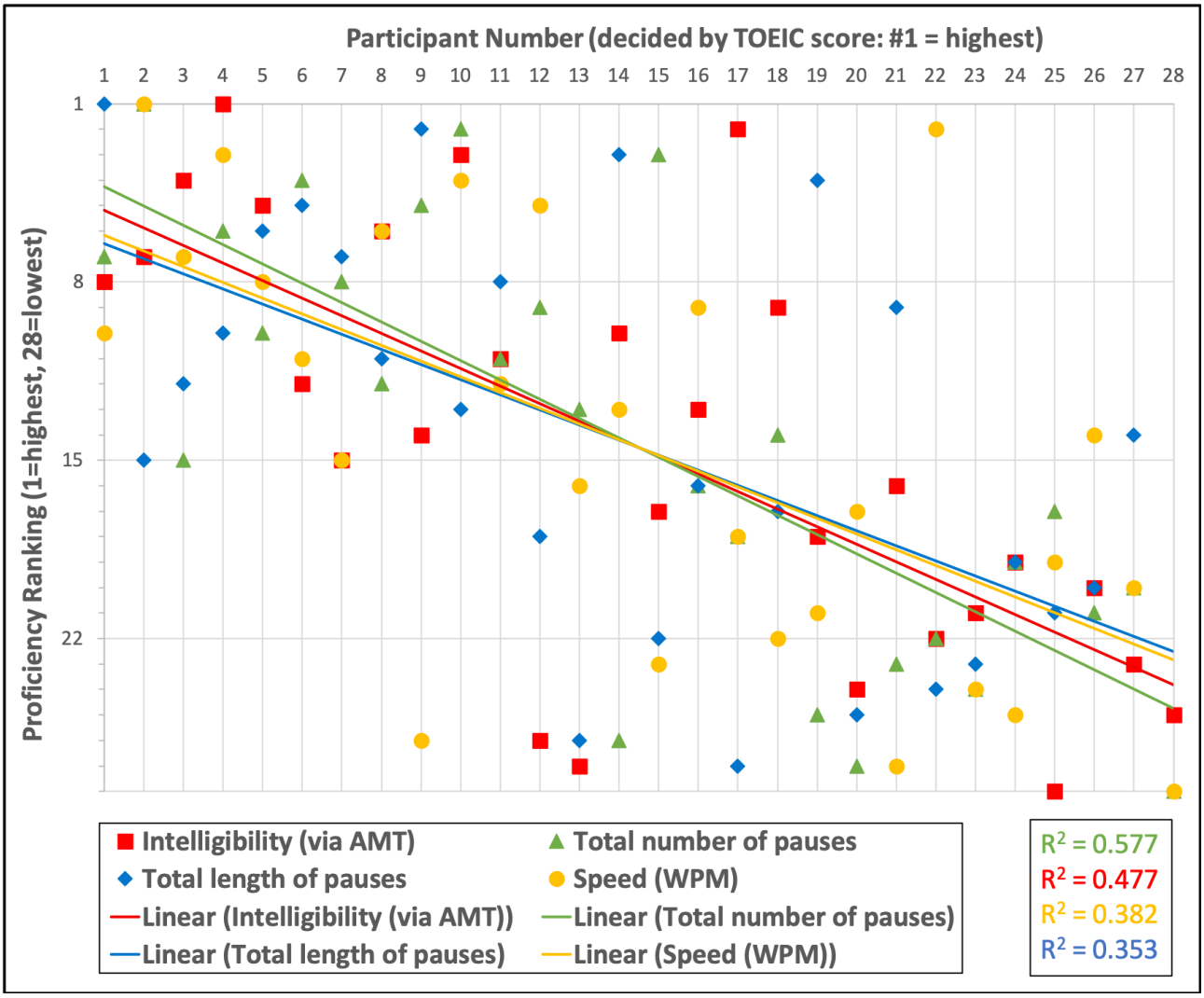

Figure 4: Proficiency ranking of all 28 speakers by each measurement method

3.4 Trial 4: Deep learning with 3-fold cross validation and flexible accuracy definition

In Trial 4, we changed only two things from Trial 3. The first one is 3-fold cross validation. Before this trial, test results were not stable because there was a possibility that features in the training data were unbalanced. To avoid this variability, we adopted the system of 3-fold cross validation in Trial 4. Cross validation is performing multiple rounds of cross-validation using different partitions and then averaging the results over all the rounds.

The second change is changing the definition of test accuracy. In the previous definition, we specified that only test data that were labeled by the computer at exactly the same proficiency level as were labeled by us were accurate. However, in this way, data labeled as even only one level out of position (e.g., proficiency group 3 participants being categorized as group 2 or 4) were considered to be hard errors (i.e., accuracy = 0) and this was not helping the computer to learn the categories as quickly and accurately. Therefore, in a new, more flexible definition of test accuracy, we set data labeled as only 1 level out of position to be soft errors ("almost correct"), and we gave those accuracies a value of 0.5 points instead of 0 points. Through this change, we could get much higher accuracy here than in our previous deep learning trials.

Figure 7 shows the highest test accuracy results for each labeling method in Trial 4. Because of the flexibility in the definition of test accuracy, almost all test accuracies are higher than in Trial 3. Labeling proficiency by total number of pauses led to the highest test accuracy (47.7%) in deep learning Trial 4.

4 Discussion

4.1 Labeling data

Usually when using supervised deep learning, training sets can be unambiguously labelled. For example, if training a computer to recognize animal images, the training data contains images of cats, dogs, elephants, etc. that are clearly labelled as such. These labels are often called the "ground truth". However, in some machine learning cases in the medical field for example, different experts diagnose (i.e., label) the same patient's problem in different ways [23]. Likewise, in our research, it is difficult to

Table 5: The 56 selected phrases from the Wolf Passage used in the AMT intelligibility test in Trials 3 & 4

No.	Phrase	No.	Phrase
1	There was once	29	that a few days later
2	poor shepherd boy	30	he tried exactly
3	who used to watch	31	the same trick again
4	his flocks	32	and once more
5	in the fields	33	he was successful
6	next to a dark forest	34	However, not long after
7	near the foot of a mountain	35	a wolf that had just escaped
8	One hot afternoon	36	from the zoo
9	he thought up a good plan	37	was looking for a change
10	to get some company	38	usual diet of chicken and duck
11	for himself	39	overcoming
12	have a little fun	40	its fear of being shot
13	Raising his fist	41	it actually did come out
14	in the air	42	from the forest
15	he ran down	43	and began to threaten the sheep
16	to the village	44	Racing down to the village
17	shouting Wolf Wolf	45	the boy of course cried out
18	As soon as	46	even louder than before
19	they heard him	47	Unfortunately
20	the villagers	48	as all the villagers
21	rushed from their homes	49	were convinced that
22	full of concern	50	he was trying to fool them
23	for his safety	51	a third time
24	and two of his cousins	52	they told him
25	even stayed with him	53	Go away
26	for a short while	54	don't bother us again
27	This gave the boy	55	And so the wolf
28	so much pleasure	56	had a feast

Table 6: Number of training and testing files for each label in Trial 3

Label	Total # of training wav files	Total # of testing wav files
Total # of pauses	29833	4943
Total length of pauses	30311	4465
WPM	29822	4954
AMT	29611	5165

establish the ground truth about someone's language proficiency, because there are so many ways of measuring it and each way may give different results.

In our research, we decided to have five levels of proficiency in Trials 1 and 2, and four levels of proficiency in Trials 3 and 4. Each level was a different label for deep learning. In Trials 1 and 2, we had expert listeners judge the nativeness of our speakers on a 7-point scale. However, this method has a problem: since the expert listeners were listening to whole sentences, they had suprasegmental information like stress and intonation, which were not available to the AMT workers, who were only judging very short phrases. It is not clear what points the expert listeners were judging nativeness on. Hence, we decided to use an objective measure – fluency indicators found in the acoustic signal.

In Trials 3 and 4, we used four methods of measuring speaking proficiency: total number of pauses, total length of pauses, words per minute, and subjective judgements by AMT participants. As seen in Figure 4, the proficiency ranking of any given participant varies somewhat depending on the labeling method, indicating that a ground truth is difficult to establish. For example, participant #1, who has the highest TOEIC score, is also ranked #1 in total length of pauses, but is ranked #7 in total number of pauses, #8 in intelligibility, and #10 in speed of speech. It may be that a combination of measurements would be a better way of labeling the proficiency of participants for the purpose of training a DNN.

In Figure 4, the straight lines are linear estimates of the rankings for each label. The R^2 values in the key under the plot show the strength of correlation between each proficiency ranking method and the TOEIC ranking of the participants. Note that the total number of pauses in the read speech for each participant correlates most strongly with their TOEIC scores (R^2=0.577). The TOEIC scores are from the Listening and Reading test and did not include Speaking or Writing, so this strong correlation makes sense. It is interesting that this same measure

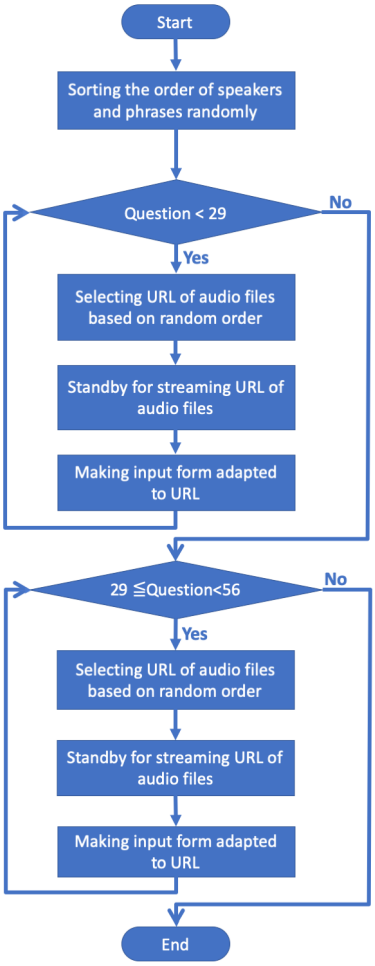

Figure 5: Flowchart of algorithm used in the HTML file for pseudo-randomization of phrases in the AMT intelligibility transcription task. This was used in Trials 3 & 4.

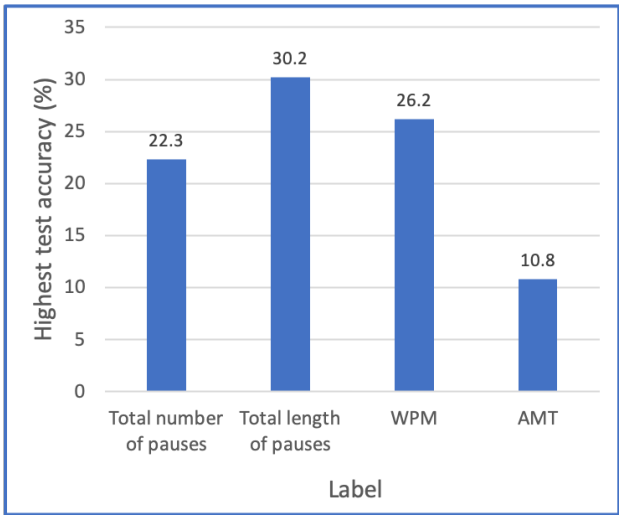

Figure 6: Deep learning highest test accuracies by labeling method in Trial 3

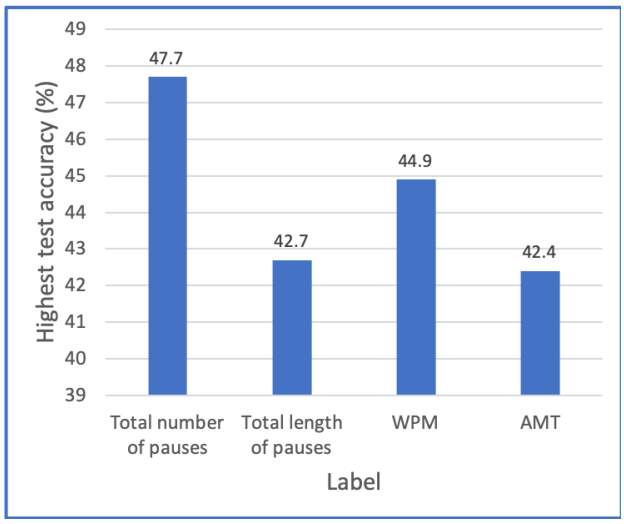

Figure 7: Deep learning highest test accuracies by labeling method in Trial 4 – with flexible definition of test accuracy

(total number of pauses) is the one that gave the highest test accuracy in the final deep learning trial (Trial 4).

As seen in the results of both Trials 3 and 4, when labeling by AMT, deep learning test accuracy was the lowest among the four ways of labeling. From these results, we can say that when a machine classifies English proficiency level from audio data, it appears to be easier for the machine to learn from speech fluency labels than from speech intelligibility measures. One reason for this phenomenon is probably due to label noise. Labeling by speech fluency (number of pauses, total length of pauses, WPM) avoids label noise because labels are calculated using objective scientific measurements. However, when labeling intelligibility, even though the calculations themselves are done objectively, the correct/incorrect transcriptions are entirely based on the subjective listening task of the AMT workers. This may have affected the accuracy of deep learning. Therefore, in the case of our audio database, the neural network can learn features more easily when data is labeled by speech fluency – specifically by the total number of pauses in the read speech of a speaker.

As a whole, the deep learning accuracy tended to be low in this study. We believe that there may be two possible reasons for this problem: overfitting in the training process and an insufficient amount of training data. Overfitting means that the production of an analysis corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably. Thus, if we prevent overfitting and also add more audio files, we may get better accuracy and would be able to use this DNN algorithm in a future system.

5 Conclusions and future work

In our research, we tested deep learning classification of speech proficiency while using several ways of labeling degree of proficiency from the data. In the case of the

labeling method, we could show it is easier for neural networks to learn elements of speech fluency rather than speech intelligibility. However, we trained and tested data by deep learning using only a fairly small database. In future work, we should test our neural network with an unknown database. Based on testing with unknown data, we could make a better system in the future.

Moreover, we labeled our data in several ways: total number of pauses, total length of pauses, WPM, and speech intelligibility from AMT. Although each of these points are important for those who are English native speakers, it does not necessarily mean that any given one of them can determine one's proficiency. If proficiency was simply decided by a single objective measurement, it would be easy for a machine to rank proficiency. However, proficiency is most probably dependent on various objective factors, and not necessarily in an equal way. In future work, it is necessary to combine various types of labeling, with each one weighted in various ways, to determine the best result of deep learning.

References

[1] B. Smith, Polyglossia **23**, 199 (2012)

[2] S. Greenberg, Acoustics Today **14**, 19 (2018)

[3] M.J. Munro, T.M. Derwing, Language and Speech **38**, 289 (1995)

[4] K. Saito, P. Trofimovich, T. Isaacs, Applied Linguistics **38**, 439 (2017)

[5] N.H. De Jong, Language Assessment Quarterly **15**, 237 (2018)

[6] Y. Iguro, Master's thesis, University of Aizu (2018)

[7] Y. Saito, *Zero kara tsukuru Deep Learning [Making deep learning from 0]* (O'Reilly Japan, Inc., Tokyo, 2016)

[8] B. Frénay, M. Verleysen, IEEE Transactions on Neural Networks and Learning Systems **25**, 845 (2014)

[9] J. Li, Y. Wong, Q. Zhao, M.S. Kankanhalli, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[10] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, L. Qu, *Making deep neural networks robust to label noise: A loss correction approach*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 1944–1952

[11] S. Moriya, Y. Yaguchi, I. Wilson, *Ultrasound tongue image denoising for comparison of first and second language speakers' tongue trajectories* (2016), poster presented at the 5th Joint Meeting of the ASA and ASJ

[12] J. Yu, K. Markov, T. Matsui, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **27**, 742 (2019)

[13] D. Deterding, Journal of the International Phonetic Association **36**, 187 (2006)

[14] P. Boersma, http://www. praat. org/ (2020)

[15] B.F. Freed, N. Segalowitz, D.P. Dewey, Studies in Second Language Acquisition **26**, 275 (2004)

[16] N. Iwashita, A. Brown, T. McNamara, S. O'Hagan, Applied Linguistics **29**, 24 (2008)

[17] T.M. Derwing, M.J. Munro, Studies in Second Language Acquisition **19**, 1 (1997)

[18] Mathworks, *Speaker identification using pitch and MFCC*, https://jp.mathworks.com/help/audio/examples/speaker-identification-using-pitch-and-mfcc.html

[19] T. Isaacs, P. Trofimovich, Studies in Second Language Acquisition **34**, 475 (2012)

[20] C. Nagle, Journal of Second Language Pronunciation **5**, 294 (2019)

[21] K. Yamato, The Journal of Foreign Language Education and Media (LET) Kansai Branch Methodology Research Group (2011)

[22] Y. Sekiguchi, *The effect of speech speed and word frequency on native listeners' comprehension of L2 speakers* (2016), Undergraduate thesis, University of Aizu

[23] H. Valizadegan, Q. Nguyen, M. Hauskrecht, Journal of Biomedical Informatics **46**, 1125 (2013)