AUTOMATIC SPEECH INTELLIGIBILITY SCORING OF HEAD AND NECK CANCER PATIENTS WITH DEEP NEURAL NETWORKS

Li Bin[1], Matthew C. Kelley[2], Daniel Aalto[3,4], Benjamin V. Tucker[2]

[1]Zhejiang University
[2]Department of Linguistics, University of Alberta
[3]Department of Communication Sciences and Disorders, University of Alberta
[4]Institute for Reconstructive Sciences in Medicine, Misericordia Community Hospital, Edmonton

libin2494@gmail.com, mckelley@ualberta.ca, aalto@ualberta.ca, benjamin.tucker@ualberta.ca

ABSTRACT

In this paper, we propose an automatic method of measuring the intelligibility of head and neck cancer patients' speech. In a retrospective chart review, speech recordings of 137 individuals treated for oral and oropharyngeal cancer were included. Recordings before the treatment and at various times after the treatment were included. Naive listeners typed the utterances, and the proportion of correctly identified words was used as the measure for intelligibility. Three different neural networks were trained and compared to predict the speech intelligibility using Mel-frequency cepstral coefficients (MFCCs), Mel-frequency filterbanks, and raw audio as input. These inputs were paired with different network architectures. The network using MFCCs as input and bidirectional long short-term memory (BLSTM) layers provided the best current performance. The model's prediction was highly correlated with the actual intelligibility score, with a linear correlation coefficient of approximately 0.69. We conclude by discussing future directions and applications for this research.

Keywords: speech production; speech intelligibility; deep neural networks; head and neck cancer

1. INTRODUCTION

For patients who have been treated for oral and oropharyngeal cancer, a major change occurs in the speech production system as a result of the treatment [9]. This change impacts the speech apparatus, and as a result, the speech intelligibility is impacted [9, 14]. Speech intelligibility is currently one of the metrics for assessing a patient's ability to communicate effectively after treatment [9, 14, 18]. One current method of assessing speech intelligibility requires a volunteer to transcribe the patients' speech and uses the accuracy of the transcription as the intelligibility score [19, 24]. In the present paper, we describe an automatic tool built to evaluate speakers' intelligibility, which would provide a more time-efficient, consistent, and objective evaluation [7, 11, 12].

The present research has two main objectives. The first objective is to create a system based on human ratings that is time-efficient and comparable to human listeners. The second is to explore whether deep neural networks, trained on a limited dataset, are able to predict intelligibility scores with little to no information about the language (top-down information). In other words, can this system predict the intelligibility score from human raters using purely acoustic input and no higher level knowledge about the language [7, 11, 12]?

2. METHODS

In this section, we describe our data and three different neural networks, which vary in terms of input format and, as a result, architecture. These models were trained and tested to assess their ability to predict human intelligibility scores.

2.1. Data

To test our hypotheses, we collected data from 137 head and neck cancer patients who had undergone speech assessment at the Institute for Reconstructive Sciences in Medicine as part of their treatment pathway. A waiver of consent was obtained, and the research procedure was approved by the Health Research Ethics Board of Alberta Cancer Committee (HREBA.CC-18-0400). A total of 335 recordings with intelligibility ratings were obtained from the hospital database using convenience sampling (availability of recordings). The sample included 189 recordings of male patients and 146 recordings of female patients. Of all the recordings, 111 were obtained before treatment onset. Treatments included surgery, radiation, chemotherapy, and prosthetic treatment (e.g., a speech prosthesis that totally occludes an opening).

During each of the recording sessions, participants read a list of 50 words and lists of sentences [19, 24]. Following the recordings, a group

of volunteers orthographically transcribed each of the recordings, resulting in a transcription of each recording by one person. From these transcriptions, an intelligibility score was calculated based on the accuracy of the transcription. Thus if a volunteer accurately transcribed 40 out of 50 words correctly the intelligibility score would be 80%. The word (sentence) intelligibility scores ranged from 12% (4%) to 100% (100%) with a median of 94 (99), mean of 86.3% (94.3%), and standard deviation of 17.9% (13.1%). Not all the patients were necessarily native English speakers, although the clinic operates in English.

We processed 335 recordings, each containing 50 words each. In total, we had 16,376 words resulting in approximately 2 hours of speech. We chose the word list recordings as they provided more variability in the intelligibility scores. Sentence scores had a larger bias in the higher accuracy range, likely because transcribers could use sentential context to reconstruct the sentence. Intelligibility scores were calculated based on the accuracy of human transcribers. Each transcriber was required to transcribe each word in the set of 50 and the intelligibility score was simply a calculation of the percent accuracy of the rater correctly recognizing the 50 words.

We identified and separated each word into individual sound files for the purpose of training and testing. The intelligibility score for the file containing each set of 50 words was assigned as the score to predict for each of the individual words.

2.2. Models

Three different models were created based on the type of input to be used in each model. The architecture of each model varied to accommodate and maximize the accuracy of the network for each type of input.

The first model made use of MFCCs as the input and a BLSTM architecture. Twelve MFCCs plus the energy term calculated on 25 ms windows spaced at 10 ms were used, plus delta and delta-delta coefficients, using the Python Speech Features library [13]. MFCCs are a classic choice for acoustic speech signal processing, and have been successfully paired with deep neural networks in tasks like automatic speech recognition and phoneme labeling [5]. MFCCs can be considered a summary of important components of the original signal, and the Mel filterbanks traditionally used in the calculation were designed with the intention of mirroring human hearing.

BLSTM layers are like regular long short-term memory (LSTM) layers in that they learn temporal dependencies in the data, but the BLSTM layers learn to use both previous and future context, as opposed to just previous context like LSTM layers. They have been used successfully in a variety of tasks involving time-series data, like phoneme labeling [4, 5].

The second model uses 40-filter Mel filterbanks with the energy term calculated over 25 ms windows spaced at 10 ms, with delta and delta-delta features, using the Python Speech Features library [13]. Its architecture was a combination of convolutional and BLSTM layers. Previous research has shown that speech recognition systems pairing Mel filterbanks with convolutional layers can achieve high recognition accuracy [16, 26, 27].

The third model uses raw audio as the input, and a combination of convolutional and LSTM layers. The raw audio was sampled at 44.1 kHz, and the first convolutional layer windows the signal based on its filter size and stride length. There are indications that networks using raw audio can perform on par with those using engineered features like MFCCs [14] and Mel filterbanks [8, 20, 25]. Networks using raw audio have also found successful results in other tasks, like phoneme labeling for forced alignment [10], speech synthesis [23], and large vocabulary continuous speech recognition [22].

The overall architecture for combining convolutional layers with recurrent layers is suggested by Chollet [2], where the massive parallelizability of convolutional layers can allow networks to train faster than if they were composed completely of recurrent layers. In our models, we expect the convolutional layers to work like filter banks after training as suggested by Palaz et al. [15]. The BLSTM layers are expected to then model temporal patterns in the features detected by the convolutional layers.

2.3. Training the networks

Before the training process began, a hold-out validation set of 10% of the words were randomly selected. The networks were trained using minibatches of 400 words. The order of the words was shuffled at each epoch. The networks were trained to minimize the mean squared error between the output and the target intelligibility score. We made use of the Adam optimizer, with the amsgrad improvement [17], but default parameters otherwise. The training process for each model was stopped after 4000 epochs. This number of epochs was selected as a compromise between computation-time and model performance. The training was carried out using Keras 2.2.2 [3] with TensorFlow 1.9.0 [1] as the backend using an NVIDIA Titan X Pascal GPU.

The best configuration thus far for the network that used MFCCs as input consisted of 4 sequential

BLSTM layers with 128 units each, followed by 1 BLSTM layer with 64 units, and ending with a fully-connected layer with 1 unit with linear activation to perform the prediction. Each BLSTM layer otherwise had default values and ReLU activation. The number of layers and units in each layer were found through manual search, whereby the hyperparameters were adjusted incrementally between training sessions based on previous results.

For the network that used the Mel filterbanks as input, the best configuration thus far has 1 two-dimensional convolution layer with 128 3x3 filters and a stride length of 1, followed by 6 convolutional layers with 256 filters of size 2x3 with a stride length of 1, then 3 BLSTM layers with 128 units each, and finished by a fully-connected layer with 1 unit with linear activation to perform the prediction. Each convolutional layer was set to pad zeros to the input so that the dimensions were not reduced due to the filter size. Note that the filters are listed with the size in the temporal dimension first, and the size in the frequency dimension second. These parameters are influenced by the architectures in [27] and [26] and tweaked using manual search. Each layer otherwise had default values and ReLU activation.

The best configuration thus far for the network that used raw audio as input consisted of the whole audio file as input, with a 1 one-dimensional convolutional layer with 128 filters of size 1000 and a stride length of 200 (producing 22.68 ms windows at 4.54 ms intervals for the convolutions). This was followed by a one-dimensional convolutional layer with 128 filters of size 5 and stride length of 1, then 5 one-dimensional convolutional layers with 64 filters of size 5 and stride length of 1. Then there were 2 one-dimensional convolutional layers with 64 filters of size 3 and stride length of 1, then 2 BLSTM layers with 128 units, followed by 1 BLSTM layer with 64 units, and ending with a fully-connected layer with 1 unit with linear activation to perform the regression prediction. Each layer otherwise had default values and ReLU activation. The number of layers and the parameters for each layer were found using manual search, as before.

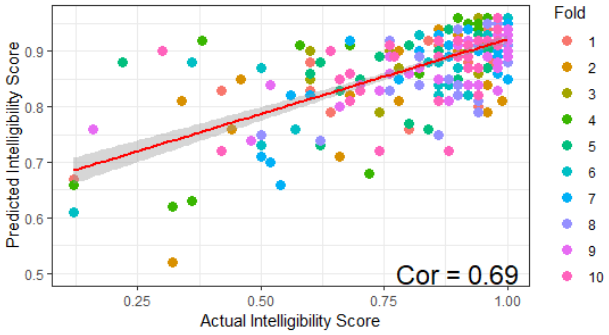
3. RESULTS

We compared each model to the others by recording the training and validation loss. The MFCC model outperformed the other two models in both training loss and validation loss, showing the highest prediction accuracy. The training loss for the raw-audio and the Mel filterbank models remained relatively unchanged even after 4000 epochs of training. For the raw-audio and fiterbank models, it seems that we have not yet found the optimal architecture for these models with the input provided. It is also possible that we have not provided them a sufficient amount of training data.

For the remainder of this paper, we focus on the MFCC model after 4000 epochs of training. We performed a 10-fold cross-validation test of the MFCC model to further validate its effectiveness. During each fold, 10% of the data set was held out for testing, and the remaining 90% was used to train the model from scratch. During this process, the data were split by participant rather than randomly. There was no overlap between the folds.

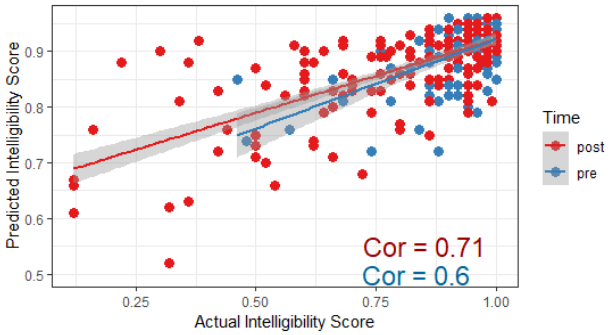
The results of the cross-validation tests are summarized in Figure 1. In the correlation plot, the predicted intelligibility scores are plotted on the y-axis and the actual intelligibility scores are plotted on the x-axis. As is clear from the correlation (0.69), MFCC model provides a good fit for the data. We further see that there is a tendency for the model to predict higher scores than are found, as there are no predicted scores below 0.5 (most are above 0.6) while there are many actual intelligibility scores below this score.

Figure 1: Correlation plot comparing predicted intelligibility scores to actual intelligibility scores . Each cross-validation fold is represented by a different color.



We also investigated the prediction accuracy of the model by splitting the data between the recordings that occurred before treatment (pre) and the recordings from after treatment (post). The results of the investigation are illustrated in Figure 2. The post-treatment recordings occurred at multiple time points after treatment, but for the purposes of this comparison, these time points have been consolidated. This consolidation is likely reflected in the fact that the correlation is lower for the pre-treatment recordings. It is clear that most of the lower intelligibility predicted values (i.e., those below 0.75) are the post-treatment values.

Figure 2: Correlation plot comparing predicted intelligibility scores to actual intelligibility scores split across pre (blue) and post (red) treatment.



Figures 1 and 2 illustrate the bias in the data for higher intelligibility scores. In other words, the training data is largely made up of scores above 0.75 and the data below that score is much sparser. The confidence interval in the lower scores also illustrates this result.

4. DISCUSSION

The present report represents our first attempt at modeling speech intelligibility using deep neural networks. We are encouraged by the initial success we have seen in this preliminary work and believe that additional work will increase the accuracy. In order to get a model with a higher degree of generalization, more data is necessary. This data would include additional accuracy scores from more than one listener along with more data from other speakers. Deep neural nets often require large amounts of training data to achieve high accuracy and good generalization. We believe that the additional data will particularly help the model in predicting scores for new input from unknown patients. The addition of this new training data is necessary for the application of this type of technology clinical situations. Further, it would be useful to have more specific intelligibility scores for individual words as the score we are currently using requires that use the average score over 50 words from one speaker and assign that value to each of the 50 words.

Another possible direction of exploration would be to investigate the intelligibility of the sentence data in addition to the word data. This has the potential to increase our training data substantially and provide training data that is more ecologically valid. We also hope to further modify the models' hyperparameters and try different architectures, which may improve the overall accuracy of the models from the different types of input.

The present work was based on head and neck cancer patients. The disease and its treatment often change the structures and articulatory control. It is possible that the network focused on characteristic acoustic changes related to changes in resonance that may predict intelligibility in this population. In addition to intelligibility, a future system could also be trained to predict variables related to social perception [18].

A final note, it would be informative to take the present set of results and use a DNN automatic speech recognition system on the patients' recordings [21] and then derive a similar speech intelligibility score as we have used already. It would be interesting to investigate similarities and differences between human scores and scores derived from the automatic speech recognition system. This modeling approach may also have the added advantage of being slightly more transparent.

5. CONCLUSIONS

The present research found that it is possible to predict accurately human intelligibility scores using deep neural networks and that MFCC based input provided the best fit to the data. This also illustrates that it is feasible to create a computational system that can predict human intelligibility scores (or recognition accuracy) with the acoustic signal and a relatively small set of training data and that no additional information about language is necessary.

6. ACKNOWLEDGMENTS

We acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. This research was funded in part by a Mitacs internship to the first author and a Canadian Social Sciences and Humanities Research Council grant to the fourth author, as well as the Kule Institute for Advanced Study through the Deep Learning for Sound Recognition group at the University of Alberta. We would like to thank the Head and Neck Surgery Functional Assessment Laboratory and Gabriela Constantinescu for facilitating the access to the data.

7. REFERENCES

[1] Abadi, M., Agarwal, M., Barham, P., Brevdo, E., Chen, Z., Citro, C., …, Zheng, X. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Available at tensorflow.org.

[2] Chollet, F. 2017. *Deep Learning with Python*. New York: Manning Publications Co.

[3] Chollet, F. 2018. Keras version 2.2.2. Available at keras.io.

[4] Graves, A., Fernández, S., Gomez, F., Schmidhuber, J. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proc. 23rd ICML* Pittsburgh, 369-376.

[5] Graves, A., Mohamed, A. R., Hinton, G. 2013. Speech recognition with deep recurrent neural networks. *Proc. 2013 IEEE ICASSP* Vancouver, 6645-6649.

[6] Graves, A., & Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18(5-6), 602-610.

[7] Haderlein, T., Moers, C., Möbius, B., Rosanowski, F., & Nöth, E. (2011). Intelligibility Rating with Automatic Speech Recognition, Prosodic, and Cepstral Evaluation. In I. Habernal & V. Matoušek (Eds.), *Text, Speech and Dialogue* (pp. 195–202). Springer Berlin Heidelberg.

[8] Hoshen, Y., Weiss, R. J., & Wilson, K. W. 2015. Speech acoustic modeling from raw multichannel waveforms. *Proc. ICASSP 2015* Brisbane, 4624-4628.

[9] Jacobi, I., van der Molen, L., Huiskens, H., van Rossum, M. A., & Hilgers, F. J. M. 2010. Voice and speech outcomes of chemoradiation for advanced head and neck cancer: a systematic review. *European Archives of Oto-Rhino-Laryngology* 267(10), 1495–1505.

[10] Kelley, M. C., & Tucker, B. V. 2018. A comparison of input types to a deep neural network-based forced aligner. *Proc. INTERSPEECH 2018* Hyderabad, 1205-1209.

[11] Kim J, Kumar N, Tsiartas A, Li M, Narayanan SS. 2015. Automatic intelligibility classification of sentence-level pathological speech. *Comput Speech Lang*. Jan; 29(1):132–44.

[12] Le, D., Licata, K., Persad, C., & Provost, E. M. 2016. Automatic Assessment of Speech Intelligibility for Individuals With Aphasia. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 2187–2199.

[13] Lyons, J. 2017. Python Speech Features. Available at github.com/jameslyons/python_speech_features.

[14] Nicoletti, G., Soutar, D. S., Jackson, M. S., Wrench, A. A., Robertson, G., & Robertson, C. 2004. Objective Assessment of Speech after Surgical Treatment for Oral Cancer: Experience from 196 Selected Cases. *Plastic and Reconstructive Surgery* 113(1), 114–125.

[15] Palaz, D., Collobert, R., Doss, M. M. 2013. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *arXiv preprint arXiv:1304.1018*.

[16] Qian, Y., Bi, M., Tan, T., & Yu, K. 2016. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(12), 2263-2276.

[17] Reddi, S., Kale, S., Kumar, S. 2018. On the convergence of Adam and beyond. *Proc. of the Sixth International Conference on Learning Representations* Vancouver, 1-23.

[18] Rieger, J., Dickson, N., Lemire, R., Bloom, K., Wolfaardt, J., Wolfaardt, U., & Seikaly, H. 2006. Social perception of speech in individuals with oropharyngeal reconstruction. *Journal of psychosocial oncology 24*(4), 33-51.

[19] Rieger, J., Wolfaardt, J., Seikaly, H., & Jha, N. 2002. Speech outcomes in patients rehabilitated with maxillary obturator prostheses after maxillectomy: A prospective study. *International Journal of Prosthodontics* 15(2), 139-144.

[20] Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., & Vinyals, O. 2015. Learning the speech front-end with raw waveform CLDNNs. In *INTERSPEECH 2015* Dresden, 1-5.

[21] Spille, C., Ewert, S. D., Kollmeier, B., & Meyer, B. T. 2018. Predicting speech intelligibility with deep neural networks. *Computer Speech & Language*, *48*, 51–66.

[22] Tüske, Z., Golik, P., Schlüter, R., & Ney, H. 2414. Acoustic modeling with deep neural networks using raw time signal for LVCSR. *INTERSPEECH 2014* Singapore, 890-894.

[23] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., Kavukcuoglu, K. 2016. WaveNet: A generative model for raw audio. a*rXiv preprint arXiv:1609.03499*.

[24] Yorkston K, Beukelman D. 1981. *Assessment of Intelligibility of Dysarthric Speech*. Portland, Ore: CC Publications.

[25] Zeghidour, N., Usunier, N., Synnaeve, G., Collobert, R., Dupoux, E. 2018. End-to-End Speech Recognition from the Raw Waveform. *Proc. INTERSPEECH 2018* Hyderabad, 781-785.

[26] Zhang, Y., Chan, W., Jaitly, N. 2017. Very deep convolutional networks for end-to-end speech recognition. *Proc. 2017 ICASSP* New Orleans, 4845-4849.

[27] Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., Courville, A. 2016. Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks. *Proc. INTERSPEECH 2016* San Francisco: 410-414.