A Comparison of GMM-HMM and DNN-HMM Based Pronunciation Verification Techniques for Use in the Assessment of Childhood Apraxia of Speech

Mostafa Shahin[1], Beena Ahmed[1], Jacqueline McKechnie[2], Kirrie Ballard[2], Ricardo Gutierrez-Osuna[3]

[1]Dept.of Electrical and Computer Engineering, Texas A&M University, Doha, Qatar
[2]Faculty of Health Sciences, The University of Sydney, Sydney, Australia
[3]Dept. of Computer Science and Engineering, Texas A&M University, College Station, Texas

10.21437/Interspeech.2014-377

Abstract

This paper introduces a pronunciation verification method to be used in an automatic assessment therapy tool of child disordered speech. The proposed method creates a phone-based search lattice that is flexible enough to cover all probable mispronunciations. This allows us to verify the correctness of the pronunciation and detect the incorrect phonemes produced by the child. We compare between two different acoustic models, the conventional GMM-HMM and the hybrid DNN-HMM. Results show that the hybrid DNN-HMM outperforms the conventional GMM-HMM for all experiments on both normal and disordered speech. The total correctness accuracy of the system at the phoneme level is above 85% when used with disordered speech.

Index Terms— Pronunciation verification, speech therapy, automatic speech recognition, computer aided pronunciation learning, deep learning

1. Introduction

Language production and speech articulation can be delayed in children due to developmental disabilities and neuromotor disorders such as childhood apraxia of speech (CAS) [1]. Traditional CAS therapy requires a child undergo extended therapy sessions with a trained speech language pathologist (SLP) in a clinic; this can be both logistically and financially prohibitive. Interactive and automatic speech monitoring tools that can be used remotely by children in their own homes, offer a practical, adaptive and cost-effective complement to face-to-face intervention sessions with a SLP.

A number of technology-based tools have been developed to facilitate general speech therapy, but a very limited number of them target the specific articulation problems of children with CAS [2], [3], [4]. The intuitive and engaging environment provided by tablets and smartphones has led to the development of generic speech therapy applications for mobile devices [5], [6]. The main drawback of all these systems is the absence of automatic feedback, which makes it hard to adapt the therapy regimen based on the specific needs of each child.

There has been limited success in incorporating automatic speech recognition (ASR) systems into speech therapy tools. This is due to the higher error rates ASR systems still exhibit for developing children due to variations in vocal tract length, formant frequency, pronunciation and grammar. Perceptual evaluations of apraxic speakers can be inconsistent and prone to error [7]. The Speech Training, Assessment, and Remediation system (STAR) [8] evaluates phoneme production by calculating the likelihood ratio produced by aligning the subject's speech using the target phoneme and alternative phonemes. In Vocaliza [9], a set of confidence measures are used to score the phoneme pronunciation level. Both systems decide whether the phoneme was pronounced correctly or incorrectly without actually detecting the errors made by the child. ASR has also been used widely in the area of second language learning. As an example, Kim et al [10], define a set of rules of the expected mispronunciations of the native Korean speakers when pronouncing an English word were defined to detect pronunciation errors. In Hafss [11], a search lattice was created from all probable pronunciation variants and fed to a speech decoder to identify errors in Quranic Arabic.

In our previous work [12], we proposed an automated therapy tool for child with CAS. The proposed system consists of 1) a clinician interface where the SLP can create and assign exercises to different children and monitor each child's progress, 2) a tablet-based mobile application which prompts the child with the assigned exercises and records their speech, and 3) a speech processing module installed on a server that receives the recorded speech, analyzes it and provides feedback to the SLP with the assessment results. The SLP can then update the exercises assigned to each child as per the feedback received. The speech processing module consists of multiple components that specialize in identifying the types of errors made by children with CAS. In [13] we presented a lexical stress classifier to detect prosodic errors.

In this paper, we enhance our earlier pronunciation verification method [12], by creating a search lattice that contains all the expected mispronunciation phonemes which includes a garbage model to collect any unexpected inserted phonemes. We also use a penalty value in both the alternative and garbage paths to control the strictness of the system. We compare the performance of two different acoustic models, the conventional GMM-HMM and the hybrid DNN-HMM [14], which has been reported to outperform the conventional GMM-HMM model in other applications [15], [16] particularly with smaller training datasets [17]. The proposed method allows us to verify the correctness of phoneme pronunciation with higher accuracy than previous pronunciation verification systems [8], [9] and provides a mechanism to detect the error type (insertion, deletion or substitution) made, if any.

The remainder of this paper is structured as follows. Section 2 describes the method and the speech corpus used. Section 3 presents the experiments performed and results. Finally, the conclusions are summarized in section 4.

2. Methods

2.1. System description

In speech therapy for CAS, the child is asked to produce a set of phonemes (utterance) based upon their prescribed regime using either a visual or oral prompt. In our automated speech therapy tool [12], the goal of the pronunciation verification module is to compare the pronunciation of each phoneme in the child's production to the prompt, and identify any mispronounced phonemes. This is achieved by creating a search lattice for each word prompt in the therapy protocol. The paths in the search lattice are based on rules of expected mispronunciations made in the children's productions as developed by a SLP after assessing 20 children with CAS.

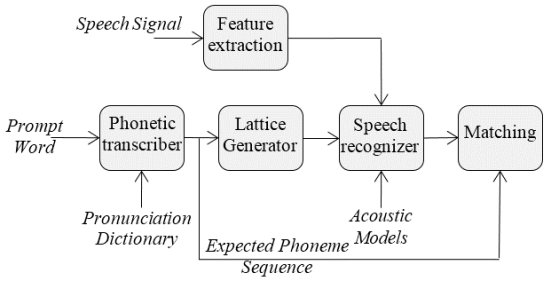

Figure 1: *Block diagram of the pronunciation verification system which uses a lattice generator and speech recognition module to compare the child's production to the given prompt.*

Figure 1 shows the block diagram of the system. The prompted word is first transcribed as per the corresponding phoneme sequence (referred to as the *correct phoneme sequence*) using the CMU pronunciation dictionary [18] and then passed to the lattice generator along with the expected mispronunciation rules. The speech signal is segmented into frames with a length of 25 msec and 15 msec overlap and a set of features extracted from each frame. The extracted features are then fed to the speech recognizer along with the created lattice and the pre-trained acoustic models to generate a sequence of phonemes from the child's utterance. An evaluation report is then generated by matching the recognized phoneme sequence with the correct phoneme sequence and specifying the errors made by the child.

2.2. Lattice creation

We used a search lattice with a specific number of alternative pronunciations for each phoneme; this limits the decoder search, making it faster and more accurate in determining the various errors made by the child. Each phoneme in the correct phoneme sequence is compared with the expected mispronunciation rules; if a rule is matched, the pronunciation variants in this rule are added as alternative arcs to the current phoneme sequence. The mispronunciation rules depend on the type of the phoneme (consonant/vowel), the phoneme position in the word (Initial/Medial/Final) and the context of the phoneme (the next phoneme in the word). Table 1 shows examples of the mispronunciation rules.

Table 1: *Examples of mispronunciation rules used*

Phoneme	Next phoneme	Position in word	Expected mispronunciations
P	Any	Initial	D
P	Any	Medial	B
K	L	Any	D/T
AE	Any	Any	AA/EY/AH

After examining all the phonemes with the mispronunciation rules, the search lattice is created using all the rules that match the correct prompt sequence. An example of a lattice created for the word "*buy*" is shown in Figure 2. A garbage model is added as an alternative to each phoneme and between phonemes to absorb any insertion of out-of-vocabulary phonemes. The garbage model consists of all the phonemes in parallel in addition to null and loop arcs to allow the decoder to either skip the garbage node (in case of no insertion) or repeat (if multiple insertions occur).

The terms PA and PG represent insertion penalties added to the alternative and the garbage arcs respectively. These penalties are added so the decoder will not align the speech to the alternative error phoneme or the garbage node unless it has enough confidence. By increasing these values, the system will tend to align the speech to the correct phonemes. These penalties can thus be used to control how strict the system is.

2.3. Acoustic models

We tested two different acoustic models, a GMM-HMM and a hybrid DNN-HMM for use in our system.

2.3.1. Conventional GMM-HMM

The system uses speaker-independent acoustic HMMs for the garbage node, which are context-independent to simplify and speed up the decoding process. However it uses tied-state context-dependent and speaker-independent HMMs for both the correct and alternative phonemes. The models are then adapted using Maximum Likelihood Linear Regression (MLLR) to produce a set of speaker-dependent models for each speaker in the test data. The features used in building the model are the Mel-frequency cepstral coefficients (MFCC). 13 coefficients are computed for each frame plus delta and acceleration to obtain a 39 dimensions feature vector per frame. The number of tied states and the number of mixtures per state are tuned using a development data set.

2.3.2. Hybrid DNN-HMM

The setup of this model is similar to the one proposed in [14]. As the DNN can easily model correlated data, Mel-scale filterbank features are used instead of the MFCCs [19]. For each frame we compute a Mel-scale filterbank with 40 coefficients and the energy together with the delta and acceleration combined into a 123 dimension feature vector. Each *n* successive frames are grouped to produce one input window to the DNN with the target label of the middle frame.

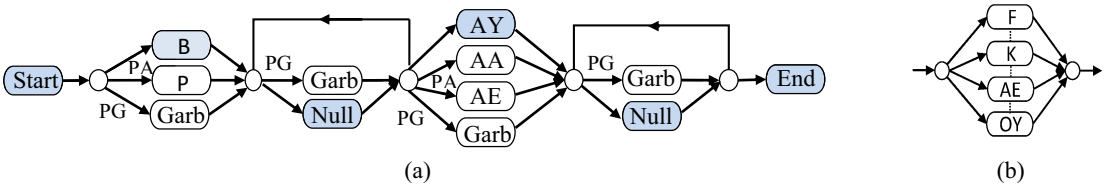

Figure 2: *(a) Lattice example of word "buy" where Garb is the garbage node. The filled nodes represent the correct phoneme sequence. (b) The construction of the garbage node.*

Here too, the number of layers of the DNN, size of each layer and number of frames in the input window are tuned with a development data set.

2.4. Speech corpus

We used two speech corpora to test our pronunciation verification system. The first speech corpus is the Oregon Graduate Institute of Science & Technology (OGI) kids' speech corpus [20]. This corpus consists of 1100 normal children from kindergarten through grade 10 saying 205 isolated words, 100 sentences and 10 numeric strings. Each utterance was verified by two individuals and classified as "good" utterances (the word is clearly intelligible with no significant background noise or extraneous speech), "questionable" utterances (intelligible but accompanied by other sounds) or "bad" utterances (unintelligible or wrong word spoken). We worked with only good utterances of the isolated words; 880 children were used to train the model, 110 children for testing and another 110 children for development.

The second speech corpus consists of disordered speech from 41 children between the ages of 4-12 years diagnosed with CAS. Inclusion criteria included no reported impairment in cognition, language comprehension, hearing or vision, orofacial structure or lower level movement programming/execution (i.e., dysarthria). Each child pronounced 90 isolated words in a speech therapy clinic in the supervision of an SLP. Each word in the data was phonetically annotated by a SLP and divided into 30, 6 and 5 speakers for training, development and testing sets respectively.

3. Experimental results

3.1. Acoustic model parameter tuning

We performed a phone decoding process using a bi-gram language model trained on the CMU pronunciation dictionary to identify the best parameter values for both the GMM-HMMs and DNN-HMMs. Decoding was applied to both speech corpora and the parameters resulting in the least phone error rate (PER) chosen for the GMM-HMMs and DNN-HMMs using the same training and development sets.

3.1.1. GMM-HMM parameter tuning

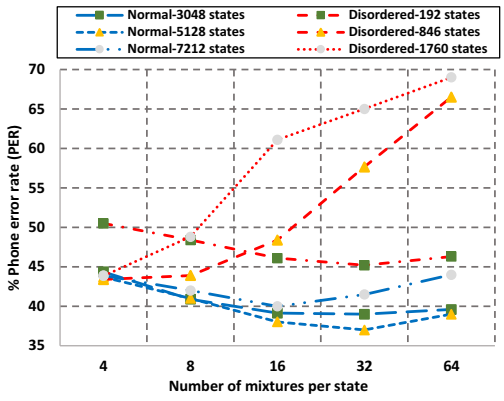

Figure 3: *Phone error rate of the development sets of both normal and disordered speech for different number of tied states and mixtures per state*

For the GMM-HMMs, we tuned the number of tied states and the number of mixtures per state. As the normal speech corpus is much larger than the disordered speech corpus, the number of tied states for the normal speech corpus was tuned from 3000 states to 7000 states while the disordered speech corpus was tuned from 200 to 1700 tied states. We evaluated different mixtures per state (4, 8, 16, 32 and 64). The PER of the development sets of both normal and disordered speech for the different parameters values is shown in Figure 3. The results show that the best PER for normal speech 37% is obtained with 5128 states and 32 mixtures per state. For disordered speech the best PER 43.4% is obtained with 846 states and 4 mixtures per state; PERs increased significantly when the number of mixtures increased as there was not enough data to train each mixture.

3.1.2. DNN-HMM parameter tuning

Next, we tested the effect of varying the number of hidden layers from 1 to 6 on a DNN-HMM with 1024 units per hidden layer and an input window of 27 frames (270 msec) to. The models were trained using training sets from both the normal and disordered corpora separately and the PER measured on development sets from both corpora. Figure 4 shows the PER of each speech corpus for different hidden layers. For normal speech, the best accuracy was obtained with 4 hidden layers, while for disordered speech the lowest PER was obtained using 2 layers; in both cases, PERs increased with additional DNN layers.

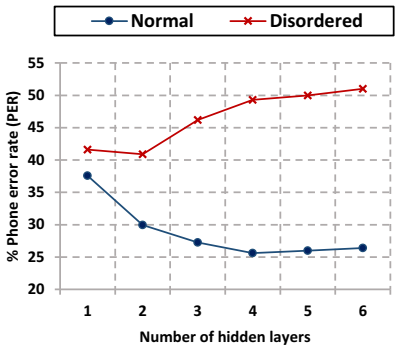

Figure 4: *Phone error rate (PER) for both normal and disordered speech corpora as a function of the number of hidden layers. The number of units in each layer is fixed to 1024 and the number of frames in the input window is fixed to 27 frames (270 msec).*

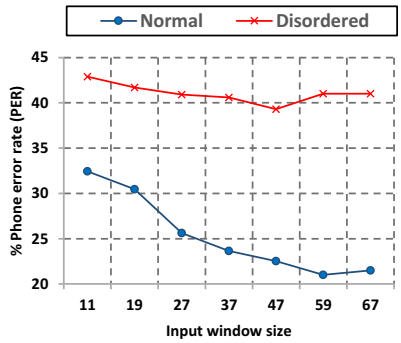

Figure 5: *Phone error rate (PER) of the development sets for both normal and disordered speech as a function of the length of the input window. The number of hidden layers is 4 for the normal speech and 2 for the disordered one and the number of units per hidden layer is 1024 for both.*

Previous work using the DNN-HMM demonstrated that increasing the input window length above 37 frames degraded the performance significantly [14]. We thus tested the effect of

varying input window length on our system performance. The PER was computed for input window sizes of 11, 19, 27, 37, 47, 59 and 67 frames with 1024 units per hidden layer and 4 hidden layers for normal and 2 layers for disordered speech. Our results in Figure 5 show that the performance kept increasing till a length of 59 frames (590 msec) and 47 frames (470 msec) for normal and disordered speech corpus respectively. Given that the words in both the training and development data of our speech corpus sets are repeated by the speakers, an increased window length is needed to develop more accurate models for each phoneme in its specific context. The resulting reduction in the generalization of the model when applied to other words, does not affect the performance of our application as the words used are limited to the words in the training set. The best PER for normal speech was 21% while for disordered speech, the best PER was 39.3%.

Given that recognition of disordered speech is a particularly challenging problem, both models performed better with normal speech than with disordered speech. The disordered speech corpus also had a reduced amount of available data for training compared to the normal speech corpus. Our results also show that the DNN-HMM performs much better than the GMM-HMM for both the normal and disordered speech corpus. The best PER obtained using the GMM-HMM was 37% and 43.4% for normal and disordered speech respectively; the PER dropped to 21% and 36.2% respectively with the DNN-HMM.

3.2. Multiple pronunciations lattice decoding

To further validate the performance of the acoustic models, we created a search lattice for each target word in the test sets of both corpora as described in section 2.4 which was then fed to a Viterbi decoder along with the extracted feature vector and pre-trained acoustic models. We used the GMM-HMMs and DNN-HMMs with the parameters that gave the best accuracy on the phone decoder as determined in Section 3.1. For normal speech we simulated pronunciation errors in correctly pronounced words by changing the labeled pronunciation sequence based on the CAS mispronunciation rules. For example, if the prompted word was "boy", the correct pronunciation sequence is "/B/ /OY/". As "B" is an alternative phoneme of "P", an error was simulated by replacing the "B" with "P" in the labeled pronunciation. The system was also tested against naturally occurring errors using the disordered speech corpus. Errors were simulated in 30% of the phonemes in each word of the normal speech corpus; and around 10% of the disordered words contained errors.

Tables 2 and 3 summarize the performance with both normal and disordered speech respectively using both acoustic models. The "correct/correct" cell represents the percentage of true positives, while the "correct/wrong" cell represents the percentage of false negatives. The percentage of false positives are listed in the "wrong/correct" cell. The percentage of the true negatives that were rejected with the same error phonemes are listed in the "wrong/wrong with same error" cell; the true negatives that were rejected with different errors from the actual error phoneme are listed in the "wrong/wrong with different error" cell. The results show that for normal speech with simulated errors, the DNN-HMM detected the mispronounced phoneme with an increase in accuracy of 4% compared to the GMM-HMM but with a small decrease in the correct acceptance. For disordered speech, the improvement with the DNN-HMM was more significant. The detection of mispronounced phonemes increased by 27.2% to 53.8% and correct phonemes by 1.9%. The total phoneme matching accuracy is around 93% for normal speech and around 89% for disordered speech. Given the absence of work on automated ASR in speech therapy for children with CAS, it is not possible to compare our resultant accuracies to existing systems. To give context to our results, in a study on the phonological difficulties of children with CAS, inter-rater reliability for the phonemic transcriptions of disordered speech by SLPs was between 78.4–97.3% [21].

Table 2: *Phoneme-level confusion matrix for* ***normal speech***

System evaluaiton	Reference evaluation				
		Correct		***Wrong***	
		GMM	*DNN*	*GMM*	*DNN*
	Correct	**96.5%**	**96%**	21.2%	16.3%
	Wrong same error	3.5%	4%	**70.1%**	**74.6%**
	Wrong different error			8.6%	9%

Table 3: *Phoneme-level confusion matrix for* ***disordered speech***

System evaluaiton	Reference evaluation				
		Correct		***Wrong***	
		GMM	*DNN*	*GMM*	*DNN*
	Correct	**91.8%**	**93.9%**	45.3%	40.5%
	Wrong same error	8.2%	6.1%	**26.6%**	**53.8%**
	Wrong different error			28.8%	5.7%

4. Conclusions

In this paper, we present a pronunciation verification method for integration into an automated speech therapy tool for children with CAS. Our approach consists of creating a search lattice for each prompt which contains the correct phoneme sequence path and a set of alternative paths that cover expected mispronunciation errors. The resulting DNN-HMM system had an overall phoneme level accuracy of 89% when used with disordered speech, which is comparable to the SLP phonetic transcription agreement of apraxic speech of around 80% [22]. Our system was able to detect the correctly pronounced phoneme with an accuracy of 94% and identify the correct errors made with an accuracy of 54%. Our proposed approach can thus be used to accurately classify mispronunciation errors in disordered children's speech collected in a noisy speech therapy environment.

We also compared the performance of conventional GMM-HMMs and hybrid DNN-HMMs. For the GMM-HMM, we achieved a minimum PER of 37% when tested with normal speech and 43% with disordered speech. When using the DNN-HMM, the PER *decreased* to 21% and 36% for normal and disordered speech respectively. For the DNN-HMM model, we demonstrated that increased window lengths are required to develop accurate phoneme models to account for the limited number of words used in speech therapy. The DNN-HMM performed better than the GMM-HMM with disordered speech, due to its ability to train better with the limited size training set. We found the DNN-HMM produced an *improvement* of 27% over the GMM-HMM in correctly detecting mispronounced phonemes and an *improvement* of about 2% in detecting correctly pronounced phonemes.

5. Acknowledgement

This work was made possible by NPRP grant # [4-638-2-236] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

6. References

[1] Adhoc Committee on CAS, ASHA. "*Childhood Apraxia of Speech*", American Speech-Language-Hearing Association, 2007.

[2] Rvachew, S., Brosseau-Lapre, F., "*Speech perception intervention*", Interventions for Speech Sound Disorders in Children, Brookes Pub, 2006.

[3] Williams, A. Multiple oppositions intervention. *Interventions for speech sound disorders in children*, Brookes Pub, 2006.

[4] Wren, Y., S. Roulstone, and A.L. Williams, "*Computer-Based Interventions*", Interventions for Speech Sound Disorders in Children. Brookes Pub, 2006.

[5] Apraxiaville. Available: http://smartyearsapps.com/

[6] *Pocket SLP*. Available: http://pocketslp.com/

[7] Kirrie J. Ballard, Donald A. Robin, Patricia McCabe, Jeannie McDonald, *"A Treatment for Dysprosody in Childhood Apraxia of Speech",* J Speech Lang Hear Res 2010;53(5):1227-1245.

[8] Bunnell, H.T., D.M. Yarrington, and J.B. Polikoff. "*STAR: articulation training for young children",* International Conference on Spoken Language Processing. 2000. 85-88.

[9] Saz, O., S. Yin, E. LLeida, R. Rose, C. Vaquero, and W.R. Rodriguez. "*Tools and Technologies for Computer-Aided Speech and Language Therapy",* Speech Communication 51 (2009): 948-967.

[10] J.-M. Kim, C. Wang, M. Peabody, and S. Seneff, "An interactive English pronunciation dictionary for Korean learners," in INTERSPEECH, 2004, pp. 1145–1148.

[11] Abdou, S.M., Hamid, S.E., Rashwan, M., Samir, A., Abd-Elhamid, O., Shahin, M., & Nazih, W., "*Computer aided pronunciation learning system using speech recognition technology*", in Interspeech, 2006.

[12] Parnandi, A., Karappa, V., Son, Y., Shahin, M., Mckechnie, J., Ballard, K., Ahmed, B., and Gutierrez-Osuna, R., "*Architecture of an Automated Therapy Tool for Childhood Apraxia of Speech*", ACM ASSETS 2013.

[13] Shahin, M., Ahmed, B. and Ballard, K., "*Automatic classification of unequal lexical stress patterns using machine learning algorithms*", Spoken Language Technology Workshop (SLT), IEEE (2012): 388-391.

[14] Mohamed, A., Dahl, G.E., Hinton, G., *"Acoustic Modeling using Deep Belief Networks"*, IEEE Trans. on Audio, Speech, and Language Processing, 2011.

[15] Dahl, G., Yu, D., Deng, L., and Acero, A., *"Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition"*, In IEEE Trans. Audio, Speech, and Language Processing, 2012.

[16] Li, L., Zhao, Y., Jiang, D., Zhang, Y., etc. *"Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition",* Proc. Conf. Affective Computing and Intelligent Interaction (ACII), pp.312-317, Sept. 2013.

[17] Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., *"Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups"*, IEEE Signal Processing Magazine, 29, November 2012

[18] *CMU pronunciation dictionary*, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[19] Mohamed, A., Hinton, G., Penn, G., *"Understanding how Deep Belief Networks perform acoustic modelling"*, in ICASSP 2012

[20] Shobaki, K., Hosom, J. P., and Cole, R. A., "*The OGI kids' speech corpus and recognizers*", Presented at the International Conference on Spoken Language Processing, Beijing, 2000.

[21] McNeill, B.C., Gillon, G.T., and Dodd, B., "Phonological awareness and early reading development in childhood apraxia of speech (CAS)", International journal of language & communication disorders / Royal College of Speech & Language Therapists, 2009, 44, (2), pp. 175-19

[22] Shriberg, L., Austin, D., Lewis, B., McSweeny, J., & Wilson, D. (1997). "*The percentage of consonants correct (PCC) metric: extensions and reliability data",* Journal of Speech, Language, And Hearing Research: JSLHR, 40(4), 708-722.