Phonemic learning based on articulatory-acoustic speech representations

Heikki Rasilo (hrasilo@ai.vub.ac.be)
Artificial Intelligence Lab, Vrije Universiteit Brussel,
Pleinlaan 9, 1050 Brussels, Belgium

Abstract

Infants learn to imitate and recognize words at an early age, but phonemic awareness develops at a later age, guided by acquisition of literacy for example. We investigate a hypothesis that speech representations in the brain are formed partly due to articulatory-acoustic learning, and these representations may be used as a basis when learning an additional mapping to phonemes. We train a convolutional recurrent neural network, having an articulatory branch and a phonemic branch for multitask learning. When trained with real conversational speech and aligned synthesized articulation, it is shown that the use of the articulatory representation boosts phoneme recognition accuracy, when the first convolutional layers are shared between the two branches. It is hypothesized that representations involved in speech perception formed in the brain during childhood may be partly based on articulatory learning, and an additional mapping from these low-level speech representations to phonemes has to be learned.

Keywords: Speech learning, speech inversion, articulatory modeling, phonetic learning.

Introduction

Despite big leaps forward in automatic speech recognition in the recent years, mainly due to the use of Deep Neural Networks (DNNs) (Hinton et al., 2012), humans still outperform machines especially under noise, or complex listening situations (Spille, Kollmeier & Meyer, 2018). Automatic speech recognition solutions are often trained to classify phones from acoustic speech. In this case the training set is segmented and annotated with phone labels. For example the widely used speech recognition dataset, TIMIT (Garofolo et al., 1993), has hand-labeled transcriptions to 61 phonetic categories. Phonetic labeling of speech is subjective by itself (Bayerl & Paul, 2011; Garofolo et al., 1993), and labeling of TIMIT is based on phonemic and allophonic knowledge of the annotators (Zue & Seneff, 1996; Keating, Byrd, Flemming & Todaka, 1994), and is thus biased by their phonemic interpretations, rather than based on purely objective speech-based information.

Categorizing continuous speech into discrete and segmental categories is a problem by itself. Indeed, research has not yet agreed on any particular universal unit of speech perception, and evidence exist that humans do not readily recognize and segment phonemes out of speech, but they have to be trained to do so (through e.g. writing). It has been shown that biasing the perception of one allophone (variant of a phoneme), does not generalize to other allophones of that phoneme (Mitterer, Scharenborg & McQueen, 2013), indicating that at least allophones can be treated as separate perceptual units in the brain (see also Reinisch, Wozny, Mitterer & Holt, 2014). Also, acquiring literacy has been shown to affect our ability to segment words into sounds (Anthony & Francis, 2015), adding to the evidence that phoneme categories have to be learned.

From a human speech learning point of view, normally developing infants learn to imitate and produce speech simultaneously with learning to recognize important speech patterns, such as words. Learning speech imitation requires learning of a mapping from acoustic speech to its continuous physical articulation, or some motor commands underlying it. Whereas this mapping is learned in early childhood, the mapping of speech acoustics into phonemes is learned at a later age, possibly having to rely on speech representations that the brain has already specialized to during the earlier learning phases. The view that speech representations are shared between the perceptual and production modalities is supported by research showing that articulatory/motor control disorders can impair speech sound perception, and this is often specific to sounds produced by the impaired articulator (see Skipper, Devlin & Lametti, 2017, for a review). The motor areas of the brain are also actively involved when listening to speech (Wilson, 2004; D'Ausilio et al., 2009)

Based on the above evidence, it seems possible that humans learn an additional mapping from context dependent sub-word units, as represented in the brain, into linguistically motivated abstract phoneme categories. From the point of view of technical solutions to speech recognition, especially phone recognition, the learning of this mapping poses an additional complication. Given acoustic speech and its phonetic transcription, the learning algorithm tries to learn a mapping between the two, but is unaware of the intermediate (perhaps articulatorily motivated) representation that humans seem to be able to conceptualize. Modern DNN solutions, given enough training data and suitable network architecture, are able to

learn complicated representations, but the amount of training data used is often much larger than what is available to a normal human language learner (e.g. 3 million 4 second utterances in 20 noise conditions, as in Sak, Senior, Rao & Beaufays, 2015). This suggests that deep neural networks, learning based on vast amounts of data, are not cognitively plausible, and thus do not increase our understanding of how humans learn to process speech.

Research has shown that grounding speech learning based on the visual modality can be used to learn robust sub-word speech units (see e.g. Harwarth, Hsu & Glass, 2019), and that learning of speech-image mappings can benefit from simultaneously learning a mapping to speech transcriptions (Chrupala, 2020). In this study we investigate if having access to an approximate representation of physical articulatory information can reduce the amount of training data needed to learn robust speech representations, that can then help to learn the phonology of a language.

Previous research

Several previous studies have investigated the use of articulatory information in speech recognition. Kirchoff (1999) described phones as discrete articulatory features (such as *voiced*, *vowel*, *front*) into which acoustic speech was mapped using a hybrid artificial neural network – hidden Markov model system (ANN/HMM). The combination of estimated articulatory features and acoustic features increased word recognition accuracy under noisy conditions. Frankel and King (2001) showed that combining measured articulatory information with acoustic features enhances speech recognition accuracy.

Mitra (2010) used the Haskins Laboratories TAsk Dynamics Application (TADA) speech synthesizer (Nam, 2004) to create synthetic speech, related trajectories for vocal tract variables (TVs, such as tongue tip position or lip aperture degree), and gestural activations. Gestures are speech action units, leading to controlled movement of tract variables, and their use is motivated by "articulatory phonology" (Browman & Goldstein, 1989), that describes speech as a series of temporally overlapping articulatory gestures. Mitra (2010) trained an articulatory gesture recognizer and showed that using TVs and acoustic features (AFs) together lead to better gesture recognition than TVs or AFs alone. He also generalizes the hypothesis to natural speech by warping synthesized trajectories over a natural speech corpus, allowing the inversion of natural speech into TVs. In word recognition experiments he shows that using inverted TVs together with AFs improves recognition over the AFs alone. Further, a Gesture-based Dynamic Bayesian Network is used for speech recognition, using AFs and inverted TVs as input, and articulatory gestures as a hidden layer. This gesture-based system provides the highest word recognition accuracy.

Mitra et al. (2017) trained neural networks to perform speech inversion based on a synthetic dataset created with TADA with several speaker characteristics. Then they recognized speech with a hybrid convolutional neural network (HCNN), using acoustic features only, or combined with inverted TVs. They report that a simple combination of features does not improve recognition accuracy when compared to the (acoustic-only) baseline, but using separate convolutional filtering on the acoustic and TV domain is needed, before combining their outputs, in order to improve recognition accuracy.

Badino, Canevari, Fadiga and Metta (2016) use datasets of acoustic speech and corresponding measured electromagnetic articulographic (EMA) data. They experiment with autoencoders to first transform raw articulatory data into a more compact representation. They train DNNs to perform acoustic-to-articulatory inversion, and report that combined acoustic and inverted articulatory features improved recognition performance when compared to acoustic features alone. Phone recognition is done using a DNN-HMM system, using DNN-based phone state classifiers. They also experiment with acoustic-to-articulatory based pre-training, where the learned inversion network is not used to provide the articulatory features, but is rather used to initialize the phone classifier. This is done by replacing the linear top layer of the network, originally providing articulatory features, into a softmax layer providing phone posteriors, and fine tuning. The pre-training technique provides a small improvement, but is not as effective as concatenation of acoustic and recovered articulatory features.

A joint model for articulatory inversion and acoustic model DNN training is introduced in Yu, Markov and Matsui (2019), showing significant improvement compared to the acoustic-only model. In their first experiment they concatenate predicted articulatory and acoustic vectors in the training and testing phases, but the articulatory inversion network is trained jointly with the rest of the network. Measured acoustic-articulatory data is used during training.

In the works described above, in the testing phase, vocal tract variables are first estimated from the speech signals to be recognized via a speech inversion system, i.e. full speech inversion from acoustics to articulation is needed in the recognition phase. There are also studies that use the articulatory representation only during training, and recognition is based on acoustic features only, and speech inversion is not needed. In the second experiment of Yu, Markov and Matsui (2019) a Generalized Distillation method is used, where a separate teacher network uses the articulatory information to learn and provide soft targets to the student network, that then learns to recognize phonemes without having to perform speech inversion. Markov, Dang and Nakamura (2006) train a hybrid HMM/Bayesian network model, where the articulatory characteristics of phoneme states are captured in the hidden variables of the Bayesian Network. They use measured articulatory data and show the hybrid network's increased recognition accuracy when compared to acoustic-only baseline. Also, Canonical Correlation Analysis (CCA) has been used to warp the acoustic representations of speech into a domain that is

motivated by articulation. CCA learns maximally correlated projections of articulatory and acoustic representations of speech (Bharadwaj, Arora, Livescu & Hasegawa-Johnson, 2012; Wang, Arora, Livescu & Bilmes, 2015; Tang, Wang & Livescu, 2018), and only the acoustic projections are used when testing. Different variants of acoustic projections discovered with CCA consistently outperform unprojected acoustic features in speech recognition. In Wang et al., (2015) a variant of CCA, where the canonical correlations are optimized using a neural network, is reported to outperform a speech inversion based system, where a mapping from acoustics to articulation is learned and the articulatory features are appended to the acoustic ones during recognition. In CCA-based studies, measured articulatory data is used.

In the current study, an important difference to previous research (except for Markov et al., 2006, the Generalized Distillation method of Yu, Markov and Matsui (2019) and the CCA studies) is that we do not use the recovered articulatory trajectories as an additional feature to acoustic speech features when learning to recognize phones. Instead, we investigate the usability of representations formed at lower layers of the neural network structure during articulatory-phonemic learning, and thus do not need to perform speech inversion during the testing phase. Mitra et al. (2014) have investigated using the hidden layers of an inversion DNN as acoustic features, but showing no improvement compared to acoustic baseline. These layers were trained based on acoustic-to-articulatory inversion alone, whereas we train these layers in the joint task of inversion and phonemic learning, showing more promising results.

Since in the current study it is not necessary to perform speech inversion during testing, our work has a similar approach to the CCA studies, and the Generalized Distillation experiment in Yu, Markov and Matsui (2019), but uses a slightly differing strategy motivated by infant speech learning. First, instead of measured articulatory data we use synthesized speech articulations that are time-aligned with a database of spontaneous, conversational, Finnish speech. The learning model thus has no access to exact articulations of the speakers, only its own vocal tract model, analogously to human learners. Second, convolutional recurrent neural networks are used to learn the mapping from acoustic speech into speech articulation and in phoneme categories at the same time. This multi-task learning model is compared with a baseline phoneme recognizer that uses the same network architecture, but learns a direct mapping from acoustics into phonemes. It is investigated at which level the articulatory branch of the network should be separated from the phonemic branch for maximal performance improvement in phoneme recognition accuracy.

Note that even though the cognitive plausibility of using DNNs with vast amounts of training data was criticized in the introduction, in this study DNNs are used as a tool to learn mappings between input and output data. This is done in order to test the potential value of taking different speech modalities into account when learning, hopefully leading to more cognitively plausible learning strategies in the future.

Experiments

In these experiments we use an articulatory speech synthesizer to create trajectories of articulatory variables over a Finnish database of conversational speech. The database used is the Aalto University DSP Course Conversation Corpus[1]. For the purpose of keeping computational time reasonable for experimentation, we use a subset of the complete dataset. We use the first 2295 speech samples for training and validation, corresponding to male speakers 1-94, and female speakers 1-12 from years 2013 and 2014. From these conversations we select the ones of less than 20 seconds in duration, to create a compact training and validation set of 2189 utterances, totaling about 3.4 hours of conversation, out of which 38 minutes has been annotated as silence.

For testing, maximally 20 second long segments of the following dataset utterances 2296 to 3000 are used, consisting of male speakers 94-116 and female speakers 13-21. The whole available dataset or the proposed division to development, training, and evaluation sets are not used, since we are interested in relative improvements on simple and reasonably rapidly training network structures, and are not aiming to reach benchmark recognition accuracies on the given data.

The dataset has automatically generated alignments to Finnish graphemes. In Finnish, each vowel phoneme, and almost every consonant phoneme, has one corresponding grapheme (Suomi et. al., 2008). Thus, the grapheme alignment is close to a phonemic transcription. The grapheme level alignment works well with LeVI articulatory text-to-speech synthesizer (see supplementary material to Rasilo et al., 2013), that can be given a sequence of Finnish phonemes and their target times, and that creates dynamic articulatory trajectories through the (intended) articulatory target positions, and synthesizes the corresponding output sound. The articulatory model creates coarticulatory effects in the way that vowel targets may not be fully reached due to new vowel or consonant target appearing close in the future. Also, consonant articulations are context dependent - consonant gestures are superposed onto existing vocal tract configuration (or movements), regarding the articulators that create the targeted consonant gesture.

For each sound file in the used subset of the speech corpus, articulatory trajectories are created by giving LeVI a phoneme target on the same time moments as the beginning of the corresponding grapheme in the data transcription. LeVI creates articulatory trajectories with a 10ms time resolution. All articulatory trajectories are normalized to have zero mean and a standard deviation of one

[1]Available at: http://urn.fi/urn:nbn:fi:lb-2015101901

Baseline experiment – speech-to-text

In the baseline experiment, a traditional speech-to-text recognizer is trained. All the experiments are conducted using Tensorflow and Keras libraries for Python. All input and output data are zero-padded to a length of 2000 frames (20 seconds with a 10ms frame shift). The output of the training set consists of one-hot encoding of the correct grapheme for each frame, following the data transcription. The graphemes included in the dataset are '#abdefghijklmnoprstuvyäö', where # corresponds to a silent frame, thus totaling 24 output categories. As an initial acoustic feature 26-dimensional log-Mel spectrograms are used, extracted with 25ms window and 10ms frame shift. The spectrograms are zero-padded to the same length as the output, and normalized to have zero mean and a standard deviation of one.

After experimenting with several CNN architectures, a network with 4 convolutional layers, 2 LSTM layers and an output layer was found to provide good recognition results for a reasonably simple model. Figure 1 (left) shows the architecture of the baseline CNN. The filter sizes for the one-dimensional convolutional layers is 8, whereas the channel size reduces when going towards the top layers of the network. One-dimensional convolutional layers are used, each kernel spanning the whole frequency range, in order to find temporal patterns of increasing complexity when moving up the network. After the convolutional layers, two stacked LSTM layers are used with tanh-activations, to capture the temporal evolution of the output of the convolutions.

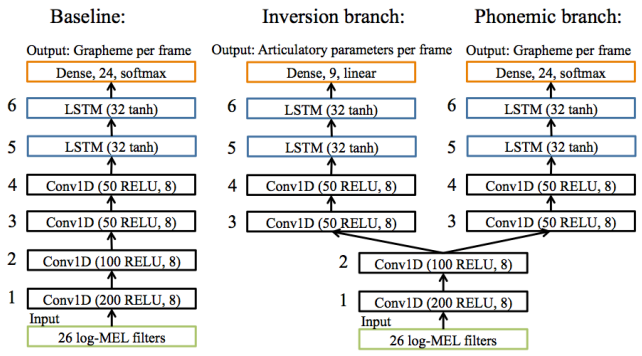

Figure 1. Used neural network architectures. Baseline network trained to recognize graphemes from speech (left). Multitask network, where training of the lowest 2 layers is influenced also by speech articulation

Multitask learning with an articulatory branch

The performance of the baseline network is compared with a branched network, where the second branch learns to perform acoustic-to-articulatory inversion, from speech acoustics into the hypothesized articulatory parameter trajectories produced by the vocal tract model. Since our initial hypothesis is that learning speech inversion can aid in finding acoustic patterns that are beneficial for phone recognition, the articulatory branch is split from the phonemic branch after a number of layers that are left common for both networks (see Fig 1, right). We investigate six different cases, having from one to six bottom layers shared by the two networks (e.g. in Figure 1, two layers are shared). The articulatory training data consists of 9 articulator position parameters that are normalized to zero mean and standard deviation of one. For every sample, a weighted sum of the losses of the two branches is used to calculate the total loss for the sample. The weights are constant over all experiments and are selected so that the two losses have approximately an equal contribution.

Training, testing and hypotheses

Each network is trained ten times, to account for the variation caused by random initialization of network weights, and the variation in the training and validation sets. For each run, 20% of the 2189 samples are randomly selected for the validation set, and the rest for the training set. Weighted categorical cross-entropy was used as the loss function for the phonemic branch – due to the highly imbalanced grapheme frequencies, weights were calculated as the inverse frequency of the grapheme counts in the training set. For the inversion branch, mean squared error loss function was used. Adam optimizer, with a learning rate of 0.0005 and a batch size of 50 samples was used. Each ten runs per network architecture were trained for 100 epochs. For each run, the network weights that result in the lowest validation loss during the 100 epochs are saved for further analysis. The ten best models per architecture are used to recognize the separate test set, and their average weighted frame classification error rate (FERs) and standard deviation is calculated. Again, FER weighting is done based on the inverse occurrence frequencies of each grapheme, giving average recognition accuracy per grapheme. Note that in the testing phase, the grapheme classification is performed only with the phonemic branch of the multitask network, requiring only the acoustic features for recognition.

The hypotheses of the several test cases are the following. We expect increased grapheme recognition performance in the multitask scenario, due to previous studies having shown that the use of articulatory information, even with synthesized articulation (Mitra et al., 2010, 2017), should increase recognition accuracy. If this is not the case, the vocal tract model may not be suitable for the purpose, or the network architecture should be better designed. If we see an increase in grapheme recognition accuracy, we hypothesize that splitting the network somewhere in the lower layers should provide optimal performance. This is due to our hypothesis that articulatory learning in early infancy probably guides learning of certain speech representations, and learning the phonemic mapping is a separate process occurring at a later age, possibly partly relying in the already learned representations.

If optimal performance is obtained splitting the network in the top layers, it indicates that inverting speech all the way to the articulatory representation may be beneficial for phone recognition. This hypothesis is additionally tested with the often-used technique of concatenating the inverted articulatory features with acoustic features. In this case, first an inversion network is trained (considering only the left branch of the multitask network in Figure 1). Then, the original acoustic features are concatenated with the inverted articulatory parameters, and a new network (with the same structure as the baseline network, except for the increased dimensionality of the input) is trained to perform grapheme recognition.

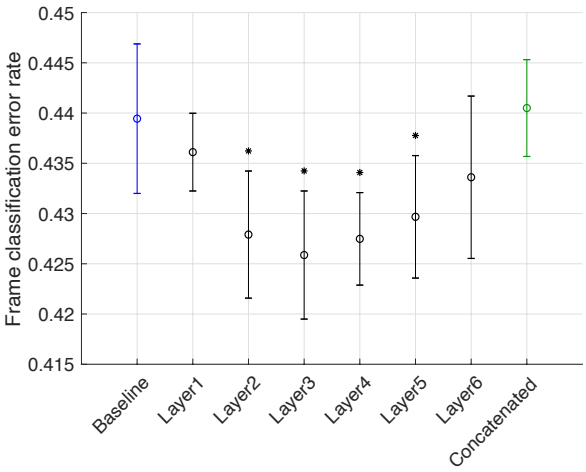

Figure 2. Frame classification errors on the test set, mean and standard deviation of the best models, after training each architecture 10 times. Baseline network, and the phonemic branch of multitask network, when split into two branches (after the layer mentioned on the x-axis) during training. Significant improvements compared to the baseline are marked with an asterisk. In the "Concatenated" model, fully inverted speech and acoustic features are concatenated for recognition.

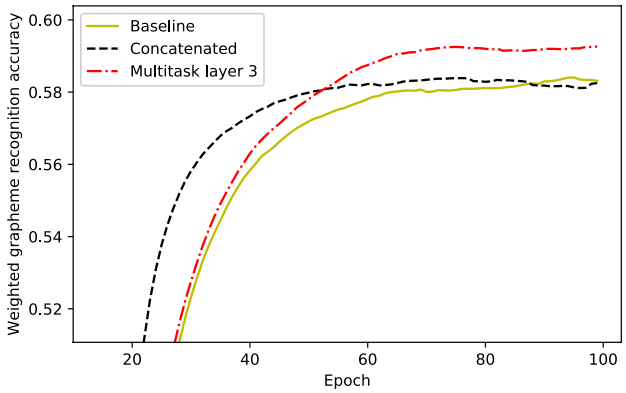

Figure 3. Weighted grapheme recognition accuracy on the validation set during training. Average over 10 runs per model, with a 20-frame moving average for smoothing.

Results

The average frame error rates of the ten best models for each architecture, and their standard deviations are drawn in Figure 2. Two-sample t-tests are performed to see if the differences between the FERs are significant ($p \leq 0.05$), when compared to the baseline case.

We see that the best grapheme recognition score is obtained when the separation to the articulatory inversion and phonemic branch is done after the third convolutional layer. This indicates that it is beneficial to use both, articulatory and acoustic, features to learn the low level representations of speech. When the separation is done above the third layer, performance slowly drops, indicating that the articulatory inversion branch and the phonemic branch have to be specialized in their own mapping tasks, on top of the shared representation. This finding is in line with the Canonical Correlation Analysis studies (e.g. Wang, et al., 2015) that show that projecting the acoustic domain into an articulatorily motivated domain improves performance more than concatenating inverted articulatory features into acoustic features. Also, Badino et al. (2016) reported that merely articulatory pre-training of the speech recognition network was not the most efficient way of using the articulatory data.

Contrary to the findings in previous studies mentioned in the introduction, in our tests concatenation of inverted articulatory features and acoustic features does not improve baseline performance. This is presumably due to inaccuracy in the speech inversion – analysis of some of the synthesized articulations shows errors where the tempo of the conversational speech is fast. Further development of the articulatory synthesizer may help to overcome these issues. However, it is interesting to see that the inversion network is still able to help in finding useful acoustic patterns in the lower layers.

Figure 3 shows the smoothed evolution of the weighted recognition accuracy of the validation set during the training for three models. The network with concatenated features learns faster than the baseline, due to the inverted

articulatory features already being informative about the grapheme category. However, it finally only reaches the baseline accuracy. The multitask network split after the third layer results in the highest accuracy.

Discussion

Even though it was hypothesized that in human learning the articulatorily motivated speech representations in the brain are formed earlier than the development of phonemic awareness, we simulated the learning of the phonemic and articulatory knowledge simultaneously. This is done in order to see the potential of shared representations in a simple simulation. In future experiments, the timing aspect could be further examined with a curriculum learning strategy, where speech is first mapped to articulation and visual/referential context, and then using the learned low-level representations to map to phonemes in a subsequent phase. In the first phase, learning a representation based only on articulation may not be sufficient nor realistic, since infants have access to a lot of referential information coming from the visual, and other sensory modalities.

In this study, a vocal tract model optimized for articulating given sequences of Finnish phonemes was used. Speech transcribed to Finnish graphemes could be used as its input due to the close relation between the phonemes and graphemes in Finnish. Any other language, transcription or vocal tract model could be used, but it is important that the vocal tract model is capable of producing realistic speech articulations based on the given transcription.

Conclusions

Our study indicates that articulatory information, synthesized on top of real speech based on its phonemic transcription, can be used along with the original speech to boost speech recognition accuracy. It is also shown that it is more beneficial to tune the low-level speech representations using the articulatory information, than to perform full speech inversion into articulatory gestures when recognizing phonemes. This finding is compatible with what we know about infant speech learning: during the first years of their lives, infants learn first to map perceived speech into articulation (learning to imitate speech; Pawlby, 1977; Jones, 2009) and general word forms, and phonemic learning occurs at a later age, influenced for example by experience with written language (Anthony & Francis, 2015). The simulations show that the phonemic network may be specialized in its own complex mapping task, and that it may build upon speech representations learned in earlier phases of speech learning.

Acknowledgements

This research was funded by Ulla Tuominen Foundation. The author would like to thank prof. Bart de Boer for valuable input to the research.

References

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current directions in psychological Science*, *14*(5), 255-259.

Badino, L., Canevari, C., Fadiga, L., & Metta, G. (2016). Integrating articulatory data in deep neural network-based acoustic modeling. *Computer Speech & Language*, *36*, 173-195.

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, *37*(4), 699-725.

Bharadwaj, S., Arora, R., Livescu, K., & Hasegawa-Johnson, M. (2012). Multiview acoustic feature learning using articulatory measurements. In *Intl. Workshop on Stat. Machine Learning for Speech Recognition*.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201-251.

Chrupała, G. (2019). *Symbolic inductive bias for visually grounded learning of spoken language*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

D'Ausilio, A., Pulvermuller, F., Salmas, P., Bufalari, I., Begliomini, C., Fadiga, L. (2009). The Motor Somatotopy of Speech Perception. Current Biology 19, 381–385.

Frankel, J., & King, S. (2001). ASR-articulatory speech recognition. In *Seventh European Conference on Speech Communication and Technology*.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, *93*.

Harwath, D., Hsu, W. N., & Glass, J. (2020). *Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech*. International Conference on Learning Representations.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82-97.

Jones, S. S. (2009). The development of imitation in infancy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2325-2335.

Keating, P. A., Byrd, D., Flemming, E., & Todaka, Y. (1994). Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication*, *14*(2), 131-142.

Kirchhoff, K., (1999). Robust speech recognition using articulatory information. Ph.D. thesis, University of Bielefeld.

Markov, K., Dang, J., & Nakamura, S. (2006). Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework. *Speech Communication*, *48*(2), 161-175.

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition, 129* (2), 356–361.

Mitra, V. (2010). *Articulatory information for robust speech recognition*. Ph.D. thesis, University of Maryland.

Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., & Saltzman, E. (2014). *Articulatory features from deep neural networks and their role in speech recognition*, IEEE international conference on acoustics, speech and signal processing (ICASSP).

Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., Saltzman, E., & Tiede, M. (2017). Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. *Speech Communication*, *89*, 103-112.

Nam, H., Goldstein, L., Saltzman, E. and Byrd, D. (2004). Tada: An enhanced, portable task dynamics model in matlab", *J. Acoust. Soc. of Am., 115*(5), 2.

Pawlby, S. (1977). *Imitative interaction. In H.R. Schaffer (Ed.), Studies in mother-infant interaction*, London: Academic Press Inc., 203–223.

Rasilo, H., Räsänen, O., Laine, U.K., 2013. Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Communication, 55*(9), 909–931.

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories?. *Journal of phonetics*, *45*, 91-105.

Sak, H., Senior, A., Rao, K., and Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. *arXiv preprint arXiv:1507.06947*.

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and language, 164*, 77-105.

Spille, C., Kollmeier, B., & Meyer, B. T. (2018). Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech & Language*, *52*, 123-140.

Suomi, K., Toivanen, J., & Ylitalo, R. (2008). Finnish sound structure. *Studia humaniora ouluensia*, *9*.

Tang, Q., Wang, W., & Livescu, K. (2018). *Acoustic feature learning using cross-domain articulatory measurements*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Yu, J., Markov, K., & Matsui, T. (2019). Articulatory and Spectrum Information Fusion Based on Deep Recurrent Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(4), 742-752.

Wang, W., Arora, R., Livescu, K., & Bilmes, J. A. (2015). *Unsupervised learning of acoustic features via deep canonical correlation analysis*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience 7*, 701–702.

Zue, V. W., & Seneff, S. (1996). Transcription and alignment of the TIMIT database. In *Recent Research Towards Advanced Man-Machine Interface Through Spoken Language*. Elsevier Science BV.