Speech Inversion with Acoustic Classification

C. Neufeld

Department of Speech-Language Pathology, Oral Dynamics Lab, University of Toronto, Toronto, Canada.
christopher.neufeld@utoronto.ca

Introduction

Inferring the shape of the vocal tract from speech acoustics is a complex machine learning problem that has been explored by many researchers over the past decades using various methodologies [1-3]. This problem is central for several domains of research. It has been argued that something akin to a `motor simulation' is an important component of human speech perception, and therefore speech inversion may be an intrinsic component of humans' ability to parse speech [4]. Modeling such behavior accurately is an important part of this line of research in human perception. It has also been demonstrated that the inclusion of articulatory data improves automatic speech recognition, and improvements in speech inversion may translate into improvements in automatic speech recognition [5]. This paper presents a novel method for improving speech-inversion with Mixture Density Networks (MDNs) [6] by classifying acoustic space into distinct aerodynamically defined classes using a Hidden Markov Model (HMM) [7], and training a separate speech inverter for each acoustic class.

Aerodynamic modes and acoustic nonlinearities

There are two types of nonlinearities which complicate the project of speech inversion. The first is related to the concept of a phase transition, where the gross characteristics of a system's output change abruptly as some control variable changes continuously. We can easily identify four basic aerodynamic phases in speech: laminar flow through the mouth during vowels, liquids and glides (here symbolized as V), laminar airflow through the nose (N), turbulent airflow during aspiration, fricatives and affricates (H), and zero airflow as during the production of a stop consonant (C). Transition between these phases may be abrupt: for instance, an infinitesimally small increase in subglottal air pressure may result in an abrupt change from laminar to turbulent airflow exiting the vocal tract. The second type of nonlinearity can be observed within phases, where large changes in the shape of the vocal tract may result in negligible spectral changes, or conversely, small changes in the shape of the vocal tract may result in significant changes [8]. These characteristics of speech acoustics greatly complicate the process of speech inversion. Here a novel approach to this problem is presented by training a set of specialist speech inverters which only estimate the vocal tract shape for some specific acoustic class. In this model, inversion consists of two steps. For some acoustic speech data, a classifier identifies which portions of the speech signal belong to which acoustic class [C, V, N, H]. Then, the trajectories of speech articulators are estimated from acoustics by specialist inverters: e.g. all those portions of the acoustic signal classified as C are inverted by a C-inverter, all those portions classified as V are inverted by a V-inverter, etc. By classifying acoustic space and training different inverters for the four classes identified here, each specialized inverter only has to learn within-phase nonlinearities, while a more general speech inverter must learn both within- and across-phase nonlinearities. This approach allows the optimization of the acoustic feature set from which articulation is derived for each acoustic class. It is found that for some acoustic classes, acoustic features derived with a relatively large window allow for the best reconstruction of articulation, while for other classes, a relatively small window size provides the best acoustic features for speech-inversion.

Data

All data comes from the TORGO database [9] which consists of speech data from video sessions, as well as 3D Electro-magnetic Articulograph (EMA) sessions, from speakers with cerebral palsy or amyotrophic lateral sclerosis, and matched controls. Here only data from matched controls is used, giving three female and four male speakers. Subjects were recorded reading English text from a screen. Stimuli include non-words, short words, restricted sentences and unrestricted sentences. The acoustic data from the video sessions was used for training and testing the acoustic classifier. The acoustic and articulatory data from the EMA sessions was used to train, validate and test the speech inverters. Acoustic data was recorded with an array microphone at 22.1 kHz, downsampled to 16 kHz, and pre-emphasized. Every 5 ms a vector of acoustic features was extracted using Hamming windows with widths ranging from 10 to 200 ms. These features were 12th order Linear Predictive Coefficients, 12th order Mel-filter Cepstral Coefficients, root-mean-squared energy, and these features' 1st and 2nd time derivatives. Articulatory data was recorded with the Carsten's AG-500 Electro-Magnetic Articulograph (EMA) at a sampling rate of 200 Hz. Lateral and vertical displacement of sensors attached to the tongue tip, body and dorsum, upper and lower lip, and jaw (TT, TB, TD, UL, LL, JA) were corrected for head movement, and filtered with an 11-point Butterworth filter with a cut-off frequency of 6 Hz. All data was normalized within subjects, and transformed by a sigmoid function of the form $1/(1+e^{-x})$ to lie on the interval [0,1].

Model outline

The speech inversion model consists of an HMM, which classifies acoustic data, and a set of MDNs, each optimized to reconstruct the trajectory of a particular articulatory channel from each acoustic class. So, for instance, in reconstructing the trajectory of the lower lip during the production of the word cancer [k^{h}ænsɹ], the HMM would ideally assign the hidden state sequence CHVNHV given the sequence of acoustic feature vectors extracted from the waveform of the word. Those portions of the acoustic signal classified as C would be inverted by an MDN trained and optimized exclusively to reconstruct the articulation of the lower lip during consonants, those portions of the acoustic signal classified as V would be inverted by an MDN trained and optimized to reconstruct articulation of the lower lip during vowels, and so forth. The construction of this model is detailed in the following sections. In outline, the data is divided into 5 sets. Audio data from video sessions is used to train the HMM, and is divided into a training set and test set. Audio and EMA data are used to construct the MDNs, and are divided into a training set, a validation set used to select the best MDN for each acoustic class, and a test set to compare the performance of this model against a baseline model. The baseline model is a set of MDNs which have been trained on the same acoustic/EMA dataset but without any acoustic classification: i.e. the baseline model is trained to reconstruct the articulation of particular EMA channels for all types of speech acoustics, rather than specific acoustic classes.

Classification

Audio speech data from the video sessions in the TORGO database was manually annotated by visual inspection of the spectrogram into four classes described above: C(onsonant), H (turbulence), N(asal) and V(owel). Four subjects (2M, 2F) contributed 100 utterances each. The remaining 3 subjects were excluded either because they did not participate in the video session, or there were significant problems with the audio recording. These

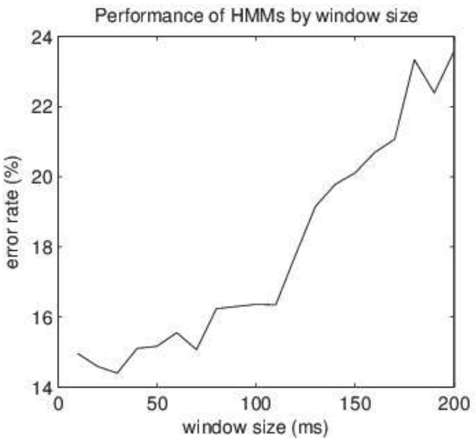

Figure 1. Accuracy of HMM as a function of acoustic window size. The best HMM is trained using acoustic features derived from a 30 ms window.

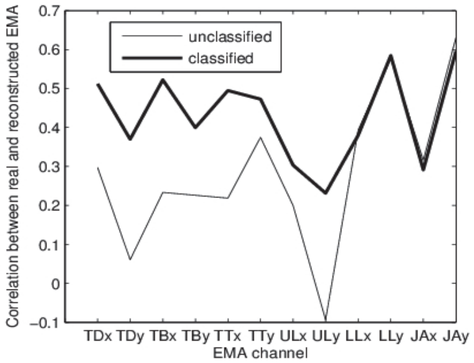

Figure 2. Correlation between real and reconstructed EMA by channel. For most channels, articulation is better reconstructed by specialist inverters operating over classified acoustic data (bold line) than by generic inverters operating over unclassified acoustic data (light

annotations were used to train a set of 4-state HMMs (each hidden state corresponds to one of the classes) with a 10-component mixture-of-Gaussians output (the acoustic features). These HMMs were trained using acoustic feature sets derived from different window sizes ranging from 10-200 ms. 90% of the data was used for training, and the remaining 10% of the data was used to test the accuracy of each HMM, and select the window size which produces the classifier with the highest accuracy. Figure 1 shows the error rate for each HMM as a function of window size. The best HMM is one where acoustic features are derived from a 30 ms window, with an overall error rate of 14.4%. Error rates for each class are [C=19.5% , V=14.4%, N=29.1%, H=10.1%].

Inversion

300 utterances each from all subjects were used for the inversion step. One subject's data was reserved for testing, and was not used for training or validation. 70% of the data from each remaining subject was used for training, 20% for validation and 10% for testing. The training data was used to construct sets of MDNs with 100 hidden units and 10 mixture components using acoustic features derived from windows ranging in size from 10 to 200 ms. For the unclassified data, the validation data was used to select the optimal acoustic window size for each EMA channel. For the classified data, validation data was used to select the optimal acoustic window size for each EMA channel *and* each acoustic class. The validation procedure for selecting the best speech inverter for the jth acoustic class and kth EMA channel: $cMDN_{j,k}$, is given here. The trajectory of the kth EMA channel estimated by the MDN trained on acoustic features derived from the ith window size for the jth acoustic class is:

1. $EMA_{i,j,k} = [MDN_i(AC_j)]_k$

C is an i x j x k matrix storing the correlation coefficient between the true articulation of the kth EMA channel underlying the acoustics of the jth acoustic class, and the estimated EMA articulation derived from the MDN trained on the jth acoustic class and ith window size

2. $C_{i,j,k} = corr(EMA_{j,k} , EMA_{i,j,k})$

The optimal MDN for the jth acoustic class and kth EMA channel is defined as that MDN for which the correlation between the true and estimated EMA is highest:

3. $cMDN_{j,k} = argmax_i(C_{j,k})$

Construction of the optimal set of MDNs for unclassified acoustics proceeds in exactly the same way, except without distinctions of acoustic class. The articulation of the kth EMA channel estimated by the MDN trained on acoustic features derived using the ith window size is:

4. $EMA_{i, k} = [MDN_i(AC)]_k$

C is an i x k matrix storing the correlation coefficient between the true articulation of the kth EMA channel, and the estimated EMA articulation derived from the MDN derived from the ith window size

5. $C_{i, k} = corr(EMA_k , EMA_{i, k})$

The best MDN for the kth EMA channel is the MDN with the highest correlation between true and estimated EMA:

6. $uMDN_k = argmax_i(C_k)$

Results

Figure 2 shows the average correlation coefficient between the real and estimated articulations for each EMA channel, for classified and unclassified speech inverters. It can be seen that, for all but the lower lip and jaw channels, classified speech inverters perform better than their unclassified counterparts. A two-way ANOVA (TYPE x CHANNEL) revealed significant main effects for both TYPE [$F(1,4776) = 11.34$, $p<10^{-16}$] and CHANNEL [$F(11,4776) = 20.21$, $p<10^{-37}$], as well as a significant interaction [$F(11,4776) = 1.64$, $p<0.0005$].

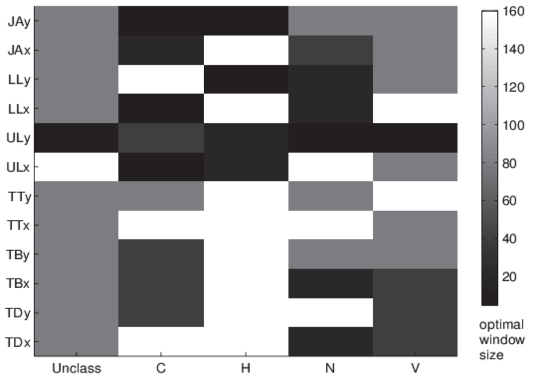

Figure 3. Window size used for each MDN by EMA channel and acoustic class. Most MDNs for unclassified data have the same window size, while the MDNs for classified data (columns C,V,N,H) employ acoustic features derived from a variety of window sizes.

Figure 3 shows the window size used to derive the acoustic features which are used for the best MDN (as determined by the validation step) for each EMA channel and acoustic class. It can be seen that, for the unclassified inverter, almost every channel has the same optimal window size, suggesting that this is an average. However, for the classified inverters, there is a very wide range of diversity, indicating that one of the means by which classified inversion achieves its superior performance is by being able to perform fine-grained optimization over specific acoustic classes and articulatory channels. For example, it can be seen that the position of articulators during turbulent airflow (class H) is generally best reconstructed using a very large window size, while for laminar oral airflow (class V) a medium-to-small window size is optimal. These results show that classifying acoustic speech data prior to inversion, and training specialist inverters to derive articulation for each acoustic class improves speech inversion. The variation of window sizes used for inverters across acoustic classes indicates that for every acoustic class, there are two levels of optimization to exploit: the window size which derives the acoustic features which are the input to the MDN, and the training of the MDN itself. This is a promising initial result, and further work is needed to explore whether these results can be replicated or improved upon with different classifier and inverter architectures.

References

1. Papcun, G., Hochberg, J., Thomas, T.R., Laroche, F., Zacks, J. and Levy, Simon. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *Journal of the Acoustical Society of America* **92**(2), 18688-18700.

2. Hogden, J., Rubin, P., McDermott, E., Katagiri, S. & Goldstein, L. (2007). Inverting mappings from smooth paths through Rn to paths through Rm: A technique applied to recovering articulation from acoustics. *Speech Communication* **49**, 361-383.

3. Rudzicz, F. (2010). Adaptive kernel canonical correlation analysis for estimation of task dynamics from acoustics. *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP'10).

4. Liberman, A. M., Cooper, F. S., Shankweiler, D. P., Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review* **74**(6), 431-461.

5. Ghosdh, P.K., Narayan, S. (2011). Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America* **130**(4): EL251-EL257.

6. Bishop, C.M. (1994). *Mixture density networks*. Technical Report. Aston University, Birmingham. (unpubl.).

7. Rabiner, L.R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* **77**(2), 257-286.

8. Stevens, K.N. 1989. On the quantal nature of speech. *Journal of Phonetics* **17**: 3-4.

9. Rudzicz, F., Namasivayam, A.K., Wolff, T. (2011). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation.* DOI 10.1007/s10579-011-9145-0