Embodied Cognition - Assignment 2

Acoustic-Articulatory Mapping and its connection with Embodied Cognition

G. Ananthakrishnan (820224-3732)

The theme of my research is the relationship between the acoustics and the articulation of speech production. This topic is handled in two main directions. The first is from articulation to acoustics, which is usually effected using either physical models or computational simulations of the physical models of the vocal-tract, the voice source and the various articulators. The data containing information about the articulation is collected from Electromagnetic Articulography (EMA), Ultra-sound, Electropalatography (EPG), MRI etc. Using such data a dynamic computational vocal-tract model is generated and the corresponding acoustics is simulated by applying the laws of acoustics [1] or by building a real mechanical vocal-tract [2]. This study is often called articulatory speech synthesis, because it uses knowledge about the way humans produce speech to synthesize it. This method is also very accurate and flexible. The collection of data is however highly tedious and expensive. Computational time required to simulate a reasonable quality of speech is very high, but the number of parameters that are required to control the production of speech are very few. When it comes to real physical models, the materials used for building the vocal tract and the method for dynamically controlling the various articulator movements (mainly tongue, jaw, glottis and the pharynx) is quite difficult.

The inverse direction of this mapping, also commonly known as inversion, is trying to predict the speech production parameters. Studying speech production from this perspective is interesting and important. Firstly this gives an idea about how human beings learn to speak. Secondly, this can form a basis of a very low-bit encoder and thirdly knowledge about the speech production parameters improves speech recognition ability significantly [3]. There are two main ways of effecting the inverse mapping, namely, inversion by synthesis [4] and statistical based inversion [5]. Inversion by synthesis makes a code-book of corresponding articulation and acoustic parameters by simulating the acoustics for different articulation parameters. Then given an acoustic parameter vector, the articulation parameters are predicted by looking them up in the code-book. Statistically based inversion uses simultaneously recorded articulatory-acoustic data to build statistical models for mapping the acoustics and articulation. Both these methods are usually quite difficult because of largely non-linear mapping as well as non-uniqueness in the acoustic-to-articulatory mapping.

Both the directions of this research are very important and relevant in trying to understand how infants learn the art of speech production. The mapping being so complex, the infants need quite some time and effort to acquire the ability to produce as well as perceive and understand adult speech. However, these are highly related as has been described in the Motor-theory of speech perception [6]. This theory proposes that humans need to perform an acoustic-to-articulatory mapping in order to classify heard speech into the different phonetic units and thus understand speech. The theory claims that by doing this, the variability within the acoustic signal under different circumstances is reduced, thus helping in speech recognition. This theory thus necessitates the presence of a human vocal-tract in order to perceive speech, which was otherwise considered a cognitive process using only perception and pattern recognition to process the acoustic features and perform a classification on them. This

theory gained support from studies which showed that the motor-cortex as well as the neurons that fire the articulatory movements are active even when humans passively read or listened to speech [7].

The process of babbling is an integral part of infants acquiring the capability to perceive and produce speech. This is the phase where the child learns the different degrees of freedom that its vocal-tract and articulators are capable of and the corresponding sounds that the vocal-tract produces. This is of course guided by speech heard by the infant from adults talking amongst themselves or infant directed speech. The central idea behind the need for babbling is that, although perceiving different sounds does not necessitate the need a production mechanism, classification and categorization of speech requires the capability to produce speech. Thus, children with perception problems seem to have speaking disabilities and vice-versa [8].

This phenomenon makes a strong candidate for studying embodied cognition, which necessitates the use of the motor-sensory connections in order to perform the cognitive task of speech recognition.

[1] Maeda S. (1988). Improved Articulatory Models. The Journal of Acoustic Society of America. 84:S146-S146.
[2] Kotaro Fukui , Toshihiro Kusano, Mukaeda Yoshikazu, Yuto Suzuki, Takanishi Atsuo, Honda Masaaki. Speech Robot Mimicking Human Articulatory Motion. In Proceeding of Interspeech. pp. 1021–1024.
[3] Wrench, A. and Richmond, K. Continuous Speech Recognition Using Articulatory Data. In Proceedings of ICSLP. pp. 145- 148.
[4] Ouni S., Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion, The Journal of Acoustic Society of America. 118(1): 444-460.
[5] Al Moubayed, S., & Ananthakrishnan, G. (2010). Acoustic-to-Articulatory Inversion based on Local Regression. In Proceedings of Interspeech pp. 937-940.
[6] Galantucci B, Fowler CA, Turvey MT. (2006). The motor theory of speech perception reviewed. Psychonomic Bulletin & Review. 13(3):361-377.
[7] Fadiga L, Craighero L, Buccino G, Rizzolatti G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eureopean Journal of Neuroscience. 15:399-402
[8] Groenen, P., Maassen, B., Crul, T., Thoonen, G. he specific relation between perception and production errors for place of articulation in developmental apraxia of speech. Journal of speech and hearing research 39(3): 468