Final Report of VSoE Research Innovation Fund (Year 2009)

Project Title: New Directions in Emotional Speech Production: Modeling of co-modulation of articulatory gestures and speech prosody

By Sungbok Lee, Shrikanth Narayanan

Background: Emotion expression through speech production is a complicated process which requires the joint control of both spatiotemporal movements of speech articulators such as the tongue, lips and jaw and speech prosody such as pitch, segmental duration and loudness. Idiosyncratic voice quality is another augmenting factor that can affect perceived emotional quality of speech stream. Traditionally speech emotion expression has been studied through the investigation of speech prosody which has long been known to the major carriers of emotional qualities. Accordingly there have been few studies on the role of articulatory maneuver in speech emotion expression and the ways the articulatory movements and prosodic control are cooperated with each other for specific emotion expressions.

We have performed some initial studies on the question of joint control of articulatory maneuver and speech prosody. Preliminary results have indicated that the articulatory movement and voice pitch are not controlled independently but jointly as a function of emotion. As illustrated in Fig. 1, we have found that there exist a tendency that pitch maneuver (e.g., pitch excursion range and pitch change rate) is more prominent for happy emotion than for angry emotion. However, articulatory activity (e.g., articulatory movement range and movement velocity) is more noticeable for angry emotion. This project has been initiated to investigate and more fully address such joint behaviors of articulatory movement and speech prosody in speech emotion expression.

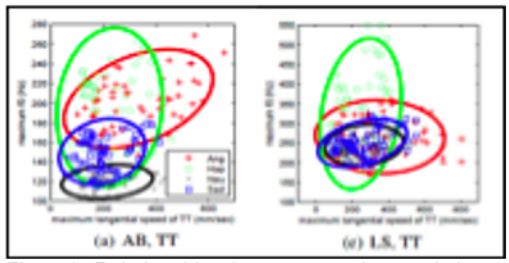

Figure1. Relationships between maximum pitch and maximum velocity in syllabic transition regions in two subjects. Differences between angry (red) and happy (green) emotions are noticeable, which implies differences in articulatory and pitch maneuvers as a function of emotion.

Aim: Since aforementioned preliminary results are limited by the number of speakers as well as by the quality of emotional speech (e.g., naïve speakers vs. actors), a primary purpose of this project has been set to collect more realistic emotional speech from professionally trained speakers such as actors and actresses with expanded speech material context.

Since perception of emotion is somewhat subjective, it's imperative to evaluate the quality of emotions expressed in speech by a large number of listeners in order to verify the target emotions expressed by speakers. This is another important purpose of this project.

Method:

(1) Speakers: 4 professional actors (2 male and 2 female) are recruited.
(2) Data collection using two hardware platforms: a 3D Electromagnetic Articulography (EMA) from the Carstens, and a 1.5T Magnetic Resonance Imaging (MRI) equipment.
(3) Speech material: 7 sentences with various linguistic contexts.

(4) Emotion types: 4 emotions (Hot-Anger, Cold-Anger, Sad, Happy) and Neutral as reference.
(5) Speaking style: 3 different speaking styles of “normal”, “loud” and “fast.”
(6) Evaluation of emotional quality: An evaluation criterion is that one utterance be evaluated by at least 5 listeners.

Current Status and Impact: As of September 2010, we have finished data collections from 4 paid professional actors (2 male and 2 female) using the EMA equipment and from the corresponding 2 female speakers using the MRI equipment. They are paid $100 for an EMA session and $50 per hour for MRI session. The matched sets of EMA and MRI vocal tract data from 2 female speakers should be useful for the study of the relation between articulatory kinematics and the whole shapes of the vocal tract as a function of emotions.

Paid-evaluation is currently on-going and two speakers’ utterances are all evaluated by a panel of 5 listeners. The other two speakers’ utterances are currently being evaluated and it will be finished until March. Evaluation MRI speech is also being planned. Each evaluator is paid $60 per subject evaluated. Evaluation is a bottle-neck on this data processing but it must be done first before any analyses.

Once evaluations are done we believe that the dataset both with EAM and MRI will provide us to explore the human emotional speech production in depth and make valuable contribution to the literature of speech production and modeling. Such results should also have impacts on the articulatory and acoustic synthesis of emotional speech. It is expected that the results of analyses of the emotional EMA and MRI databases should provide us good materials including conference and journal papers that can be further realized as successful grant application(s). Currently a conference paper is being prepared for submission to ICSLP-Interspeech 2011, and an NIH proposal is also being prepared for submission before the spring break of the 2011 academic year.