# Quantitative Analysis of Relations between Acoustic and Articulatory Speech Features

*Hani Camille Yehia*

Universidade Federal de Minas Gerais

hani@ufmg.br

The speech production process can be understood as the result of articulatory gestures generating speech acoustics. In this talk we will analyze, quantitatively, physical relations between articulatory and acoustic speech features (Fig. 1). This analysis is necessary to assess to which extent articulatory gestures can be effectively inferred from speech acoustics, taking into account physical and morphological constraints (Yehia and Itakura, 1996). This degree of inference can then be used to distinguish information that can be physically determined from information that requires prior knowledge based on speech interaction to be acquired. After a brief description of how vocal tract geometry determines speech acoustics, a quantitative analysis of measured vocal tract motion, facial and head motion, and speech acoustics is presented (Fig. 2) (Yehia et al., 1998). It is verified that, in addition to the physical coupling, there exists a functional coupling which allows the perceiver to infer acoustic features of speech from visual information (e.g. correlation between head motion and F0 contour). The results obtained are then illustrated by means of facial animation driven by speech acoustics and of speech synthesis driven by facial and head motion (Yehia et al., 2002). The motion generated in the animation, in spite of not being identical to the original, is perceptually realistic. In turn, the speech synthesis generated is intelligible, but not natural. Finally, we present new results which take into account the time-varying coupling between articulatory and acoustic speech features (Fig. 3). This makes a significant difference from previous analyses, since perceivers can handle time fluctuations present in the coupling between articulatory and acoustic speech features, which are not captured by static mappings (Barbosa et al., 2012). Also, the time-varying patterns observed deserve a deeper analysis. Possibilities for future investigation are presented at the end.

**Figure 1:** Examples of relations between articulatory and acoustic speech features: the power spectrum envelope is directly related to the vocal-tract transfer function; and the vocal-tract cross-section area function can be inferred from the speech signal or calculated from the vocal-tract geometry. (After Yehia and Itakura, 1996)

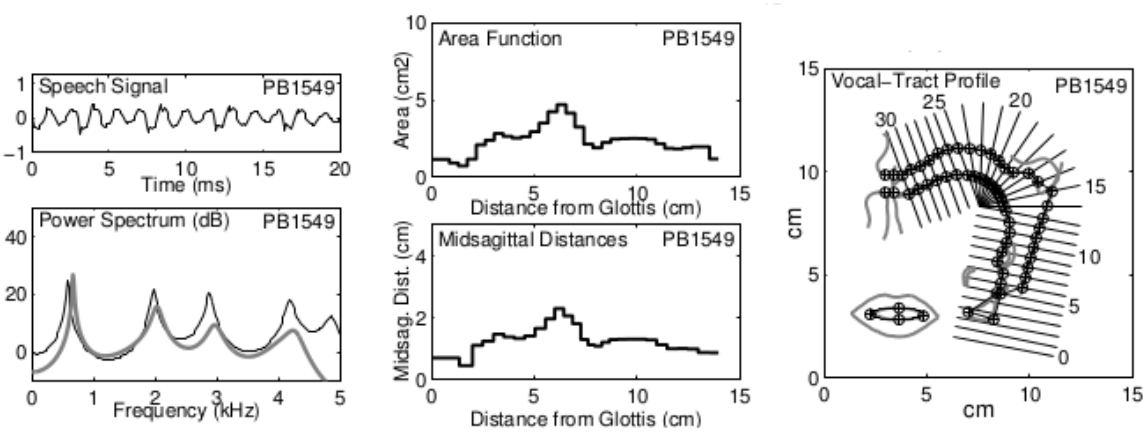


**Figure 2:** Solid lines: black – the portion of total facial variance accounted for as a function of the number of vocal-tract components used to represent the face; gray – the portion of total vocal-tract variance accounted for as a function of the number of facial components used to represent the vocal tract. Dashed lines: - the portion of vocal-tract (black) and facial (gray) variances accounted for as a function of the number of principal components used to represent vocal tract and facial motion of two subjects. (Source: Yehia et al., 1998)

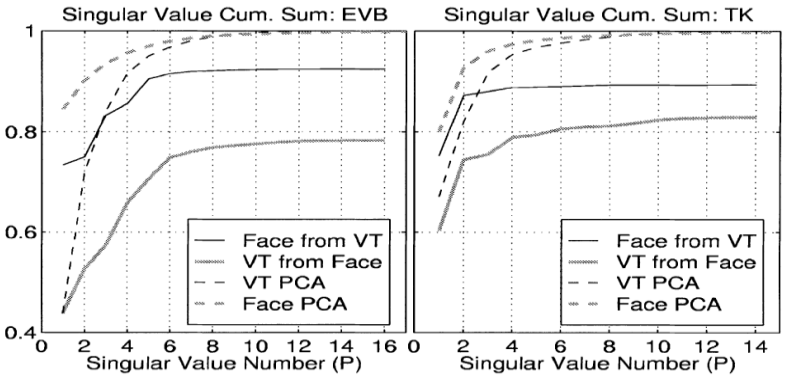
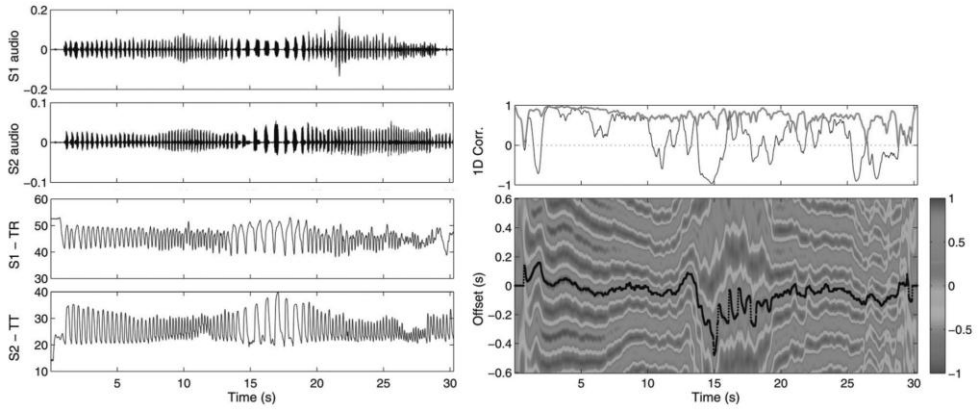



**Figure 3:** Simultaneous speech. First and second panels: speaker S1 says "cop" repeatedly, while speaker S2 repeats "top." Third and fourth panels: tongue root (TR) position of speaker S1 and tongue tip (TT) position of speaker S2. Fifth panel: 1D correlation signals for lag zero (black line) and for the optimum correlation path (red line), obtained by tracking the maximum correlation values around lag zero in the correlation map. Bottom panel: correlation map between the TR and TT signals, with the optimum correlation path shown by the black line. The color bar indicates the sign (color) and value (shade) of the correlation. (Source: Barbosa et al., 2012)




# References


Yehia, H. C., Itakura, F. 1996. A method to combine acoustical and morphological constraints in the speech production inverse problem. Speech Communication 18(2), 151-174.

Yehia, H. C., Rubin, P., Vatikiotis-Bateson, E. 1998. Quantitative association of vocal-tract and facial behavior. Speech Communication 26(1-2), 23-43.

Yehia, H. C., Kuratate, T., Vatikiotis-Bateson, E. 2002. Linking facial animation, head motion and speech acoustics. Journal of Phonetics 30(3), 555-568.

Barbosa, A. V., Dechaine, R.-M., Vatikiotis-Bateson, E., Yehia, H. C. 2012. Quantifying time-varying coordination of multimodal speech signals using correlation map analysis, Journal of the Acoustical Society of America 131(3), 2162-2172.


***