Speech Analysis using Physical Models of the Human Vocal Tract

Takayuki Arai (Sophia University, Japan)

Usually speech analysis is directly done on speech sounds uttered by humans. However, there is an alternative method, known as analysis-by-synthesis (e.g., Bell, *et al.*, 1961), which analyzes speech sounds using models. In analysis-by-synthesis, a computational model with a certain set of parameters synthesizes an output signal, and the output is compared with the target speech sound for the analysis. If the output sound has similar acoustic characteristics, then it is likely the model properly represents the target sound. If necessary, the parameters are modified to more closely approximate the target. This method can also be extended using physical models.

Arai (2012b) describes a study in which physical models of the human vocal tract are used for speech analysis. In this study, we developed physical models of the vowels /a/ and /i/ with a nasal cavity. Using these models, we tested the degree of nasality of vowels by opening the velopharyngeal port during vowel production. For results, additional poles and zeros were observed on the transfer function with acoustic coupling to the nasal cavity as a side branch to the main vocal tract (e.g., Fujimura, 1960, 1961; Fujimura & Lindqvist, 1971). The spectrum of the output sound was affected by the additional poles and zeros in the following manner: there was a reduction in amplitude of the first formant (F1) and a shifting of the frequency of F1 (e.g., Fujimura, 1960; Fujimura & Lindqvist, 1971; Maeda, 1993). These acoustic cues are consistent with predictions made by acoustic theory for nasalization.

Similar to the models developed for nasalized vowels, we have recently developed a physical model for approximants (Arai, 2013, 2014). With this model, the first half of the tongue can be rotated and raised to produce English /l/ and /r/. It is often difficult to measure the actual human vocal tract during speech production, so we conducted impulse response measurements with the /r/ model for different tongue rotations. A swept-sine signal was used for the measurements, and each response signal was converted to an impulse response. This technique is conceptually the same as the old measurement with a sweep-tone signal used by Fujimura and Lindqvist (1971). By looking at a spectral representation from the impulse response data, we see that the third formant (F3) descends below 2000 Hz in the middle of /r/, which is the main acoustic correlate of this sound.

Thus, these examples demonstrate that physical models of the human vocal tract can be used for speech analysis, and that they offer advantages over using either the actual human vocal tract or computer-based models. Some of the advantages are that physical models are 1) noninvasive, as actual human bodies are not used, 2) highly reproducibile, and 3) intuitive.

Currently, we are conducting additional experiments with physical models. First, we are looking at how breathiness of a voice source affects the perceived degree of nasalization (Arai, 2006). Because breathiness and nasalization have some acoustic cues in common (Ohala and Amador, 1981; Arai, 2006), it is often difficult to analyze nasalized vowels with breathy voice. Physical models can greatly simplify this problem because the nasalization and breathy features can be isolated. Secondly, we are looking at the "saturation effect" (Fujimura and Kakita, 1979) during production of bunched /r/ (Arai, 2014). Fujimura and Kakita (1979) pointed out that the quantal effect (e.g., Stevens, 1972) between biomechanics and acoustics during vowel production yielded stable formants in frequency. This quantal effect by Stevens (1972) was also observed in Arai's three-tube model (Arai, 2012a). Arai (2014) also pointed out the stability for bunched /r/ while testing physical models. Further discussions are under investigation.

Acknowledgments

I acknowledge Professor Osamu Fujimura for his kind encouragement regarding my vocal-tract models. This work was partially supported by JSPS KAKENHI Grant Number 24501063.

References

Arai, T. (2006). "Cue parsing between nasality and breathiness in speech perception," *Acoustical Science and Technology*, 27(5), 298-301.

Arai, T. (2012a). "Education in acoustics and speech science using vocal-tract models," *J. Acoust. Soc. Am.*, 131(3), Pt. 2, 2444-2454.

Arai, T. (2012b). "Acoustic analysis of formant shifts in nasalized vowels," *the Phonetician*, 104-105, 2012-I-II , 39-50.

Arai, T. (2013). "Physical models of the vocal tract with a flapping tongue for flap and liquid sounds," *Proc. of the Interspeech*, 2019-2023.

Arai, T. (2014). "Retroflex and bunched English /r/ with physical models of the human vocal tract," *Proc. of the Interspeech*.

Bell, C. G., Fujisaki, H., Heinz, J. M., Stevens, K. N. and House. A. S. (1961). "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Am.*, 33, 1725.

Fujimura, O. (1960). "Spectra of nasalized vowels," *Res. Lab. Electron. Q. Prog. Rep.*, No. 58, MIT, 214-218.

Fujimura, O. (1961). "Analysis of nasalized vowels," *Res. Lab. Electron. Q. Prog. Rep.*, No. 62, MIT, 191-192.

Fujimura, O., and Lindqvist, J. (1971). "Sweep-tone measurements of vocal-tract characteristics," *J. Acoust. Soc. Am.*, 49, 541-558.

Fujimura, O. and Kakita, Y. (1979). "Remarks on the quantitative description of the lingual articulation," in *Frontier in Speech Communication Research*, B. Lindblom and S. Öhman, Eds. (Academic Press, London, U.K.), 17-24.

Maeda, S. (1993). "Acoustics of vowel nasalization and articulatory shifts in French nasal vowels," in *Nasals, Nasalization, and the Velum*, M. K. Huffman and R. A. Krakow, Eds. (Academic Press, San Diego, CA), 147-167.

Ohala, J. J. and Amador, M. (1981). "Spontaneous nasalization," *J. Acoust. Soc. Am.*, 69, S54-S55.

Stevens, K. N. (1972). "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human communication: A unified view*, P. B. Denes and E. E. David Jr., Eds. (McGraw Hill, New York), 51-66.