Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 989–1010

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Phrase-level speech simulation with an airway modulation model of speech production[☆]

Brad H. Story [*]

Speech Acoustics Laboratory, Dept. of Speech, Language, and Hearing Sciences, University of Arizona, 1131 E. 2nd St., P.O. Box 210071, Tucson, AZ 85721, United States

Received 28 June 2012; received in revised form 12 October 2012; accepted 15 October 2012
Available online 26 October 2012

Abstract

Artificial talkers and speech synthesis systems have long been used as a means of understanding both speech production and speech perception. The development of an airway modulation model is described that simulates the time-varying changes of the glottis and vocal tract, as well as acoustic wave propagation, during speech production. The result is a type of artificial talker that can be used to study various aspects of how sound is generated by humans and how that sound is perceived by a listener. The primary components of the model are introduced and simulation of words and phrases are demonstrated.

Keywords: Vocal tract; Vocal folds; Modulation; Speech simulation; Speech synthesis

1. Introduction

Speech is produced by transforming the motion of anatomical structures into an acoustic wave embedded with the distinctive characteristics of speech. This transformation can be conceived as a modulation of the human airway system on multiple time scales. For example, the rapid vibration of the vocal folds modulates the airspace between them (i.e., the glottis) on the order of 100–400 cycles per second to generate a train of flow pulses that excites the acoustic resonances of the trachea, vocal tract, and nasal passages. Simultaneous, but much slower movements of the tongue, jaw, lips, velum, and larynx can be executed to modulate the shape of the pharyngeal and oral cavities, coupling to the nasal system, and space between the vocal folds by adduction and abduction maneuvers. These relatively slow modulations shift the acoustic resonances up or down in frequency and valve the flow of air through the system, thus altering the characteristics of the radiated acoustic wave over time, and providing the stimulus from which listeners can extract phonetic information.

The view that human speech is produced by a modulation system was expressed by Dudley (1940) in an article called "The carrier nature of speech." In it he referred to the relatively high-frequency excitation provided by phonation or noise generation as "carrier waves" that are modulated by slowly-varying, and otherwise inaudible, movements of the vocal tract called "message waves." He based this view on experience in developing both the VOCODER (Dudley,

[☆] This paper has been recommended for acceptance by Bernd Moebius.

[*] Tel.: +1 520 626 9528; fax: +1 520 621 9901.
E-mail address: bstory@email.arizona.edu

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.10.005

1939) and the human-operated VODER (Dudley et al., 1939), and, in the conclusion, made a curious point that a wide variety of carrier signals – even nonhuman sounds such as instrumental music – could be modulated by the "message waves" and still produce intelligible "speech." This points to the importance of understanding articulatory movement in terms of how it modulates the shape of pharyngeal and oral airspaces over time which, in turn, modulates the acoustic characteristics of the speech signal. Traunmüller (1994) also proposed a modulation theory in which speech signals are considered to be the result of articulatory gestures, common across speakers, that modulate a "carrier" signal unique to the speaker. In this theory, however, the carrier signal is not simply the excitation signal, but includes any aspects of the system that are phonetically neutral and descriptive of the "personal quality" of the speaker. This suggests that embedded within the carrier would be contributions of the biological structure of the vocal tract as well as any idiosyncratic vocal tract shaping patterns, all of which would be modulated during speech production by linguistically meaningful gestures.

Studying speech as a modulation system can be aided by models that allow for enough control of relevant parameters to generate speech or speech-like sounds. Within such models, the shape of the trachea, vocal tract, and nasal passages is usually represented as a tubular system, quantified by a set of *area functions* (cf., Fant, 1960; Baer et al., 1991; Story et al., 1996). This permits computing the acoustic wave propagation through the system with one-dimensional methods in the time domain (cf., Kelly and Lochbaum, 1962; Maeda, 1982; Strube, 1982; Liljencrants, 1985; Smith, 1992) or frequency domain (Sondhi and Schroeter, 1987). Typically for speech, only the vocal tract portion, along with the nasal coupling region, is considered to be time-varying. Thus, the challenge for developing a model that can "speak" is to define a set of parameters that allow efficient, time-dependent control of the shape of the vocal tract area function and coupling to the nasal system.

An articulatory synthesizer is perhaps the most intuitively-appealing approach to controlling the vocal tract because the model parameters consist of positions and movements of the tongue, jaw, lips, velum, etc. These are often represented in the two-dimensional midsagittal plane (cf., Lindblom and Sundberg, 1971; Mermelstein, 1973; Coker, 1976; Maeda, 1990; Scully, 1990) or as more complex three-dimensional models of articulatory structures (Dang and Honda, 2004; Birkholz et al., 2006, 2007), where, in either case, articulatory motion can be simulated by specifying the temporal variation of the model parameters. At any given instant of time, however, the articulatory configuration must be converted to an area function by empirically-based rules in order to calculate acoustic wave propagation, and ultimately produce an acoustic speech signal suitable for analysis or listening (e.g., Rubin et al., 1981; Birkholz et al., 2010; Bauer et al., 2010).

Other approaches consist of parameterizing the area function directly, rather than attending specifically to the anatomical structures. These are particularly useful when precise control of the vocal tract shape is desired. Early examples of this approach are the three-parameter models of Fant (1960) and Stevens and House (1955) in which the area function was described by a parabola controlled by a primary constriction location, cross-sectional area at that location, and a ratio of lip opening length to its area. These models were later modified to include various enhancements (Atal et al., 1978; Lin, 1990; Fant, 1992, 2001). Another type of area function model was proposed by Mrayati et al. (1988). The model parameters were not directly related to articulation but rather to portions of the area function determined to be acoustically-sensitive to changes in cross-sectional area.

Using any of these types of models to produce connected, coarticulated speech requires that the parameters allow for blending of the vowel and consonant contributions to the vocal tract shape. Ohman (1966, 1967) suggested that a consonant gesture (localized constriction) is superimposed on an underlying vowel substrate, rather than considering consonant and vowel to be separate, linearly sequenced gestures. Based on this view, Båvegård (1995), Fant and Båvegård (1997) detailed an area function model in which the vowel contribution was represented by a three-parameter model (Fant, 1992), as mentioned previously, and a consonant constriction function that could be superimposed on the vowel configuration to alter the shape of the area function at a particular location. Ohman's notion of vowel and consonant overlap has also influenced theoretical views of speech motor control. Gracco (1992), for instance, suggested that the vocal tract be considered the smallest unit of functional behavior for speech production, and that the movements of the vocal tract could be classified into "shaping" and "valving" actions. Relatively slow changes of overall vocal tract geometry coincide with the *shaping* category and would generally be associated with vowel production, whereas *valving* actions would impose and release localized constrictions, primarily for consonants.

Story (2005a) introduced an area function model conceptually similar to that of Båvegård (1995), Fant and Båvegård (1997). That is, the model operates under the assumption that consonantal, or more accurately, obstruent-like constrictions can be superimposed on an underlying vowel-like area function to momentarily produce an occlusion or partial