Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 49 (2007) 186–200

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Voiced speech as response of a self-consistent fundamental drive

Friedhelm R. Drepper

Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

Received 28 March 2006; received in revised form 15 January 2007; accepted 15 January 2007

Abstract

Voiced segments of speech are assumed to be composed of non-stationary acoustic objects which can be described as stationary response of a non-stationary fundamental drive (FD) process and which are furthermore suited to reconstruct the hidden FD by using a voice adapted (self-consistent) part-tone decomposition of the speech signal. The universality and robustness of human pitch perception encourage the reconstruction of a band-limited FD in the frequency range of the pitch. The self-consistent decomposition of voiced continuants generates several part-tones which can piecewise be confirmed to be topologically equivalent to corresponding acoustic modes of the excitation on the transmitter side. As topologically equivalent image of a glottal master oscillator, the self-consistent FD is suited to serve as low frequency part of the basic time-scale separation of auditive perception and to describe the broadband voiced excitation as entrained (synchronized) and/or modulated primary response. Being guided by the acoustic correlates of pitch and loudness perception, the time-scale separation avoids the conventional assumption of stationary excitation and represents the basic decoding step of an advanced precision transmission protocol of self-consistent (voiced) acoustic objects. The present study is focussed on the adaptation of the trajectories (contours) of the centre filter frequency of the part-tones to the chirp of the glottal master oscillator.

Keywords: Signal analysis; Non-stationary acoustic objects; Part-tone phases; Fundamental drive; Cascaded response; Generalized synchronization; Voiced continuants

1. Introduction

For decreasing signal to noise ratio, human speech perception shows an outstandingly lower increase of the error rate of word recognition, when compared to automatic speech recognition (ASR) (Lippmann, 1997; Gold and Morgan, 2000). The higher robustness of human acoustic perception has been demonstrated in many different situations of speech communication including context free presentation and the cocktail party. In the latter situation the use of frequency modulated voiced speech turns out to generate more robust communication than the use of unmodulated voiced speech (McAdams, 1989) and in particular than the use of whispered speech. When combined with the higher sensitivity of ASR and of whispered speech to the distance and direction of the acoustic communication, these empirical findings open an unconventional perspective on the reasons of the well known problems of ASR. These problems may be caused to a lesser extent by the obvious imperfection of the linguistic or statistical language models and to a larger extent by an insufficiently differentiated description of the aero-acoustics of voiced speech.

In spite of the undisputedly high degree of non-stationarity of speech signals, the present day determination of acoustic feature vectors of ASR is based on the assumption that speech production can be described as a linear time invariant (LTI) system (on the time scale of about 20 ms). The wide sense stationarity of an LTI system is typically used either as prerequisite for the consistent estimation of Fourier spectra or of autoregressive (all pole) models (Gold and Morgan, 2000; Vary et al., 1998; Schroeder, 1999). In the latter case it is common practice to introduce a drive–response (input–output or source–filter) model, which restricts the stationary autoregressive

E-mail address: f.drepper@fz-juelich.de

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.01.004

description to the resonance properties of the vocal tract. Linear autoregressive models are suited to describe transients with varying decay rates in different frequency ranges including relatively long (resonant formant) transients. The average decay rates (Liapunov exponents) of such transients are known to represent topological invariants (Kantz and Schreiber, 1997) (which can be assumed to be invariant under changes of the geometry of the acoustic transmission) and important cues for the distinction of vowels (Gold and Morgan, 2000; Vary et al., 1998; Schroeder, 1999). However, the conventional LTI system approach turns out to be problematic in the case of voiced phones. The vocal tract filter should not be assumed to be time invariant (Gold and Morgan, 2000; Vary et al., 1998; Schroeder, 1999) and the source not to be generated by an autonomous *linear* dynamical system (Teager and Teager, 1990; Jackson and Shadle, 2001; Herzel et al., 1994; Kubin, 1995).

Low-dimensional autonomous *nonlinear* dynamical systems have been introduced to describe newborn infant cries and dysphonic adult voices (Herzel et al., 1994) and have also been found to bring additional accuracy to models of normophonic voices (Kubin, 1995). However, there is empirical evidence that the complex neural control of the vocal fold dynamics leading to shimmer, jitter and vocal tremor (Schoentgen, 2001) impedes or precludes a low-dimensional *autonomous* deterministic description of the phonation process. Being hopelessly irregular from the point of view of acoustics, the time evolution of pitch and loudness (intonation and prosody) can partially be given phonological interpretation (Grice, 2006). The connection to linguistics and para-linguistics (emotions) invalidates or challenges a *stationary* stochastic process description of the voice source.

A more differentiated and physiologically plausible phenomenological description of the aero-acoustics of voiced speech can be achieved by introducing an additional drive–response step, which describes the highly complex wideband acoustic source as stationary (primary) response of a *non-stationary*, band-limited fundamental drive process in the frequency range of the pitch (Drepper, 2004, 2005a,b,c, 2006). The importance, generality and precision of the acoustic percept of pitch can be taken as a first hint that the hidden fundamental drive (FD) can directly be extracted from the speech signal. This leads to a two-level cascaded drive–response model (DR model) of voiced speech production which describes the speech signal as secondary response of a hidden FD. The two levels of the response cannot only be interpreted (more or less erroneously) as source and vocal tract filter output but can also be used with advantage to introduce two complementary types of simplification of the cascaded response dynamics.

In case of the secondary response it is common practice to simplify the vocal tract resonances by assuming time invariant stable *linear* response dynamics (resulting from an all pole filter) with a fixed point attractor. (An attractor is an invariant set of states which homes the asymptotic long time behaviour of the dynamics. Stable linear dynamical systems have a single trivial point attractor with dimension $d = 0$, the origin of state space.) In case of the primary response there is no doubt that voiced continuants (sustainable phones with active phonation) are generated by *nonlinear* dynamics with attractors of dimension $d > 0$ (Herzel et al., 1994; Kubin, 1995; Schoentgen, 2001). Complementary to the long transients of a typical (vowel type) secondary response, the primary response is assumed to result from strongly dissipative nonlinear dynamics (Kantz and Schreiber, 1997), which generates predominantly short transients. Such dynamics can be simplified drastically by restricting the dynamics to the asymptotic invariant set (which neglects the transient behaviour).

Invariant sets (attractors) with dimension $d > 0$ of *autonomous* (nonlinear) deterministic dynamical systems are known to represent either continuous manifolds (limit cycles or tori) or fractal sets (homing chaotic dynamics) (Kantz and Schreiber, 1997), whereas *unidirectionally coupled* (DR) systems with dissipative responses are known to have invariant sets which are subsets of continuous synchronization manifolds (lines or surfaces) in the combined state space of drive and response (Afraimovich et al., 1986; Rulkov et al., 1995). Being constrained to a continuous synchronization manifold, the (primary) response can be expressed by a continuous coupling function which describes the momentary state of the response by a unique function of a response related state of the (fundamental) drive. Synchronization or phase-locking is known to be a generic property of nonlinearly coupled DR dynamics (Kantz and Schreiber, 1997). As a rather general form of synchronization of band-limited oscillators with different frequencies, phase locking (synchronization of phases) can be detected by choosing an oscillator description in terms of amplitude and phase variables and by restricting the synchronization analysis to the phases of the oscillators (Drepper, 2000).

As will be explained in more detail, the distinction between the acoustic source and the FD opens the option to reconstruct a coherent hidden drive for a complete voiced speech segment. The latter feature can be used to reveal additional features of voiced speech, which are invariant (robust) under changes of the acoustic communication channel and to separate the phonetically relevant fast dynamics from intonation and prosody without invoking the assumption of stationary excitation.

The idea that the higher frequency acoustic modes of voiced speech, song and music as well as the perception of their pitch are causally connected to a single acoustic mode in the frequency range of the pitch (*son fundamentale* or fundamental bass), can be traced back to Rameau (1737). However, Seebeck (1844) could show that (virtual) pitch perception does not rely on a fundamental acoustic mode which is part of the heard signal. (In the meantime Fourier had invented an efficient method to separate *periodic* modes with different frequencies.) Seebeck was also the first to use a time periodic "impulse function" which