Available online at www.sciencedirect.com
SciVerse ScienceDirect
Computer Speech and Language xxx (2013) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise☆

Tuomo Raitio[a,*], Antti Suni[b], Martti Vainio[b], Paavo Alku[a]

[a] *Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*
[b] *Department Behavioural Sciences, University of Helsinki, Helsinki, Finland*

Received 14 October 2012; received in revised form 23 January 2013; accepted 23 March 2013

Abstract

This papers studies the synthesis of speech over a wide vocal effort continuum and its perception in the presence of noise. Three types of speech are recorded and studied along the continuum: breathy, normal, and Lombard speech. Corresponding synthetic voices are created by training and adapting the statistical parametric speech synthesis system GlottHMM. Natural and synthetic speech along the continuum is assessed in listening tests that evaluate the intelligibility, quality, and suitability of speech in three different realistic multichannel noise conditions: silence, moderate street noise, and extreme street noise. The evaluation results show that the synthesized voices with varying vocal effort are rated similarly to their natural counterparts both in terms of intelligibility and suitability.

Keywords: Statistical parametric speech synthesis; Adaptation; Vocal effort; Lombard speech; Breathy speech; Intelligibility

1. Introduction

Humans adapt their vocal communication to the acoustic and auditory environment in order to successfully and efficiently deliver a message to a listener without using unnecessary effort. In environments with high levels of interfering noise, more effort is required in order to increase the signal-to-noise ratio (SNR) and thereby the intelligibility of speech. This automatic effect is known as the Lombard effect, and the speech produced in such conditions is called Lombard speech (Lombard, 1911) or speech-in-noise (Langner and Black, 2005). When speaking in silence or in low noise conditions, such an effort is not necessary, and the use of a softer voice is considered more appropriate. Thus, depending on the context, natural speech varies greatly from whispering or soft phonation to shouting. This variation in the use of vocal effort is called the vocal effort continuum.

Even though the vocal effort continuum is an integral part of human communication, it is typically not utilized in machine-to-human communication. In order to produce contextually appropriate synthetic speech, the auditory environment and context must be taken into account and speech must be produced at the corresponding point in the vocal effort continuum. The modeling of the vocal effort continuum in speech synthesis not only increases the

☆ This paper has been recommended for acceptance by R.K. Moore.
* Corresponding author. Tel.: +358 50 4410733; fax: +358 9 460224.
E-mail address: tuomo.raitio@aalto.fi (T. Raitio).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.03.003

intelligibility of the system in adverse conditions, but also makes the synthetic voice more natural and more appropriate for the listener (Raitio et al., 2011b). Thus, a message delivered through such a text-to-speech (TTS) system is more likely to be comprehended by the listener.

Conventionally, the modeling of the vocal effort continuum has been difficult in TTS due to the limitations of the techniques used. In concatenative speech synthesis (Hunt and Black, 1996; Black and Campbell, 1995), large amount of speech data is required to cover sufficient units along the continuum. However, with statistical parametric speech synthesis techniques (Zen et al., 2009), the construction of such a continuum is relatively easy by using adaptation techniques (Yamagishi et al., 2009). One or a few smaller databases recorded along the continuum can be used to adapt the statistical parametric voice to any point on the continuum.

GlottHMM (Raitio et al., 2011c) is a statistical speech synthesis system that parametrizes speech by modeling the functioning of the real human speech production mechanism. Recently, GlottHMM was shown to enable synthesizing rather natural sounding and highly intelligible normal and Lombard speech (Raitio et al., 2011b). In the current study, work on GlottHMM is extended by creating a continuum from low to high vocal effort using three databases along the continuum: breathy, normal, and Lombard speech. The intelligibility and contextual appropriateness are evaluated by synthesizing both female and male speech on the continuum in three realistic multichannel noise conditions: silence, moderate street noise, and extreme street noise. Compared to the authors' previous work on Lombard speech synthesis (e.g. Raitio et al., 2011b), the present investigation encompasses also the study of breathy speech, thus extending the vocal effort continuum to soft phonation. In addition, a more advanced synthesis technique is utilized, and the scope of the present study is also expanded by including female speech and a more extensive and versatile subjective evaluation.

The paper is organized as follows. The background of the study is discussed in Section 2 by addressing how vocal effort is reflected in different speech parameters and by describing previous investigations in the study area. Section 3 gives a detailed description of the method used in this study for creating synthetic voices along the vocal effort continuum. Subjective evaluation of the voices in realistic noise environment and the consequent findings are described in Section 4. Finally, Section 5 discusses the relevance and implications of the results and summarizes the paper.

2. Background

2.1. Properties of vocal effort

The acoustic properties of speech sounds change not only as a function of linguistic message, but also according to speaker, context, and expression. This study concentrates on the use and reproduction of different levels of vocal effort, which can depend on linguistic (Gordon and Ladefoged, 2001) and all of the three extralinguistic properties (Gobl and Ní Chasaide, 2003) mentioned above. Change in the vocal effort can be triggered by a noisy environment, in which case it is called the Lombard reflex or Lombard effect (Junqua, 1993). Change in the vocal effort can also be triggered by the need to communicate over a distance (Traunmüller and Eriksson, 2000) or a change in emotional expression (Ishi et al., 2010; Gobl and Ní Chasaide, 2003). This study will mainly address the case where the noise environment is the cause for the vocal effort changes, but also the effect of distance is discussed.

The effects of vocal effort on speech has been widely studied (see e.g. Rostolland, 1982; Summers et al., 1988; Junqua, 1993; Traunmüller and Eriksson, 2000). In high vocal effort, word duration is reported to be longer, the vowel duration increased, and the consonant duration generally decreased compared to normal speech. The mean fundamental frequency (f_0) of speech is also increased and its variance is decreased in high vocal effort. The formant frequencies are shifted due to the more open vocal tract. Especially the first formant frequency (F_1) increases and the bandwidth decreases while the second formant frequency (F_2) may decrease. The spectral emphasis of speech is also shifted from low frequencies in low vocal effort to mid or high frequencies in high vocal effort. High vocal effect is characterized by decreased spectral tilt, which is due to the increased subglottal pressure and increased vocal fold tension, thus creating a more abrupt closure of the vocal folds. Finally, one of the most prominent consequences caused by increased vocal effort is the rising of the sound pressure level (SPL) of speech, increasing the SNR and thus intelligibility of speech. However, it is important to note that the vocal effort is a subjective phenomenon, different from the purely objective quantitative measure represented by SPL. The effects of increased or decreased vocal effort are generally similar, but, as Junqua (1993) emphasizes, the effects of vocal effort may vary according to speaker. As reported by Summers et al. (1988), the effects of increased vocal effort are also related to those of clear speech.

Please cite this article in press as: Raitio, T., et al., Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.003