Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 442–450
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate

Adriana Stan [b,*], Junichi Yamagishi [a], Simon King [a], Matthew Aylett [c]

[a] The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, UK
[b] Communications Department, Technical University of Cluj-Napoca, 26-28 George Baritiu St., 400027 Cluj-Napoca, Romania
[c] CereProc Ltd., Appleton Tower, 11 Crichton Street, Edinburgh EH8 9LE, UK

Received 13 July 2010; received in revised form 6 December 2010; accepted 6 December 2010
Available online 17 December 2010

Abstract

This paper first introduces a newly-recorded high quality Romanian speech corpus designed for speech synthesis, called "RSS", along with Romanian front-end text processing modules and HMM-based synthetic voices built from the corpus. All of these are now freely available for academic use in order to promote Romanian speech technology research. The RSS corpus comprises 3500 training sentences and 500 test sentences uttered by a female speaker and was recorded using multiple microphones at 96 kHz sampling frequency in a hemianechoic chamber. The details of the new Romanian text processor we have developed are also given.

Using the database, we then revisit some basic configuration choices of speech synthesis, such as waveform sampling frequency and auditory frequency warping scale, with the aim of improving speaker similarity, which is an acknowledged weakness of current HMM-based speech synthesisers. As we demonstrate using perceptual tests, these configuration choices can make substantial differences to the quality of the synthetic speech. Contrary to common practice in automatic speech recognition, higher waveform sampling frequencies can offer enhanced feature extraction and improved speaker similarity for HMM-based speech synthesis.

Keywords: Speech synthesis; HTS; Romanian; HMMs; Sampling frequency; Auditory scale

1. Introduction

Romanian is an Indo-European Romance language and has similarities with Italian, French and Spanish. Due to foreign occupation and population migration through the course of history, influences of various languages such as those of the Slavic family, Greek and Hungarian can be found in the Romanian language.

Currently, there are very few Romanian text-to-speech (TTS) systems: Most systems are still based on diphones (Ferencz, 1997) and the quality is relatively poor. To the best of our knowledge, only Ivona provides commercially-acceptable good quality Romanian synthesis; it is based on unit selection (Black and Cambpbell, 1995; Hunt and Black, 1996).[1] For promoting Romanian speech technology research, especially in speech synthesis, it is therefore essential to improve the available infrastructure, including free large-scale speech databases and text-processing front-end modules.

With this goal in mind, we first introduce a newly recorded high-quality Romanian speech corpus called "RSS",[2] then we describe our Romanian front-end modules and the speech synthesis voices we have built.

* Corresponding author.
E-mail addresses: adriana.stan@com.utcluj.ro (A. Stan), jyamagis@staffmail.ed.ac.uk (J. Yamagishi), simon.king@ed.ac.uk (S. King), matthew@cereproc.com (M. Aylett).

[1] See respectively http://tcts.fpms.ac.be/synthesis/mbrola.html, http://www.baum.ro/index.php?language=ro&pagina=ttsonline, and http://www.ivona.com for Romanian diphone system provided by the MBROLA project, Baum Engineering TTS system, Ancutza, and Ivona unit selection system.

[2] Available at http://octopus.utcluj.ro:56337/RORelease/.

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.12.002

HMM-based statistical parametric speech synthesis (Zen et al., 2009) has been widely studied and has now become a mainstream method for text-to-speech. The HMM-based speech synthesis system HTS (Zen et al., 2007c) is the principal framework that enables application of this method to new languages; we used it to develop these Romanian voices. It has the ability to generate natural-sounding synthetic speech and, in recent years, some HMM-based speech synthesis systems have reached performance levels comparable to state-of-the-art unit selection systems (Karaiskos et al., 2008) in terms of naturalness and intelligibility. However, relatively poor perceived "speaker similarity" remains one of the most common shortcomings of such systems (Yamagishi et al., 2008a).

Therefore, in the later part of this paper, we attempt to address this shortcoming, and present the results of experiments on the new RSS corpus. One possible reason that HMM-based synthetic speech sounds less like the original speaker than a concatenative system built from the same data may be the use of a vocoder, which can cause buzziness or other processing artefacts. Another reason may be that the statistical modelling itself can lead to a muffled sound, presumably due to the process of averaging many short-term spectra, which removes important detail.

In addition to these intrinsic reasons, we hypothesize that there are also extrinsic problems: some basic configuration choices in HMM synthesis have been simply taken from different fields such as speech coding, automatic speech recognition (ASR) and unit selection synthesis. For instance, 16 kHz is generally regarded as a sufficiently high waveform sampling frequency for speech recognition and synthesis because speech at this sampling frequency is intelligible to human listeners.

However speech waveforms sampled at 16 kHz still sound slightly muffled when compared to higher sampling frequencies. HMM synthesis has already demonstrated levels of intelligibility indistinguishable from natural speech (Karaiskos et al., 2008), but high-quality TTS needs also to achieve naturalness and speaker similarity.[3]

We revisited these apparently basic issues in order to discover whether current configurations are satisfactory, especially with regard to speaker similarity. As the sampling frequency increases, the differences between different auditory frequency scales such as the Mel and Bark scales (Zwicker and Scharf, 1965) implemented using a first-order all-pass function become greater. Therefore we also included a variety of different auditory scales in our experiments.

We report the results of Blizzard-style listening tests (Karaiskos et al., 2008) used to evaluate HMM-based speech synthesis using higher sampling frequencies as well as standard unit selection voices built from this corpus. The results suggest that a higher sampling frequency can have a substantial effect on HMM-based speech synthesis.

The article is organised as follows. Sections 2 and 3 give details of the RSS corpus and the Romanian front-end modules built using the Cerevoice system. In Section 4, the training procedures of the HMM-based voices using higher sampling frequencies are shown and then Section 5 presents the results of the Blizzard-style listening tests. Section 6 summarises our findings and suggests future work.

2. The Romanian speech synthesis (RSS) corpus

The Romanian speech synthesis (RSS) corpus was recorded in a hemianechoic chamber (anechoic walls and ceiling; floor partially anechoic) at the University of Edinburgh. Since the effect of microphone characteristics on HTS voices is still unknown, we used three high quality studio microphones: a Neumann u89i (large diaphragm condenser), a Sennheiser MKH 800 (small diaphragm condenser with very wide bandwidth) and a DPA 4035 (headset-mounted condenser). Fig. 1 shows the studio setup. All recordings were made at 96 kHz sampling frequency and 24 bits per sample, then downsampled to 48 kHz sampling frequency. This is a so-called over-sampling method for noise reduction. Since we oversample by a factor of 4 relative to the Nyquist rate (24 kHz) and downsample to 48 kHz, the signal-to-noise-ratio improves by a factor of 4. For recording, downsampling and bit rate conversion, we used ProTools HD hardware and software.

The speaker used for the recording is a native Romanian young female, the first author of this paper. We conducted 8 sessions over the course of a month, recording about 500 sentences in each session. At the start of each session, the speaker listened to a previously recorded sample, in order to attain a similar voice quality and intonation.

Fig. 1. Studio setup for recordings. Left microphone is a Sennheiser MKH 800 and the right one is a Neumann u89i. The headset has a DPA 4035 microphone mounted on it.

[3] Another practical, but equally important, factor is footprint. In unit selection, higher sampling frequencies may lead to a larger footprint. However, the use of higher sampling frequencies does not in itself change the footprint of a HMM-based speech synthesis system. The use of higher sampling frequencies increases computational costs for both methods.