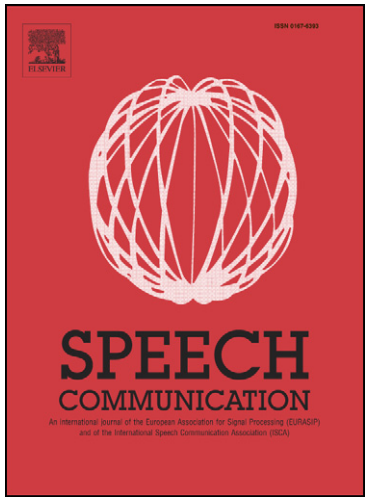
The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate

Adriana Stan, Junichi Yamagishi, Simon King, Matthew Aylett



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

Adriana Stan, Junichi Yamagishi, Simon King, Matthew Aylett

PII: S0167-6393(10)00207-4
DOI: 10.1016/j.specom.2010.12.002
Reference: SPECOM 1954

To appear in: *Speech Communication*

Received Date: 13 July 2010
Revised Date: 6 December 2010
Accepted Date: 6 December 2010

Please cite this article as: Stan, A., Yamagishi, J., King, S., Aylett, M., The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate, *Speech Communication* (2010), doi: 10.1016/j.specom.2010.12.002

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

Adriana Stan[b,1], 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*

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 96kHz 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 practise in automatic speech recognition, higher waveform sampling frequencies can offer enhanced feature extraction and improved speaker similarity for HMM-based speech synthesis.

Key words: speech synthesis, HTS, Romanian, HMMs, sampling frequency, auditory scale

1. Introduction

Romanian is an Indo-European Romance language and has similarities to Italian, French and Spanish. Due to foreign occupation and population migration through the course of history, influences of various languages such as Slavic, Greek, 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)[2]. 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"[3], then we describe our Romanian front-end modules and the speech synthesis voices we have built.

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

Email address: Adriana.Stan@com.utcluj.ro (Adriana Stan)
[1] Corresponding author

[2] 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.

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

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.[4]

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

Figure 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.

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.

The recording scripts comprised newspaper articles, sentences from novels, two short fairy tales written by the Romanian author Ion Creangă, and semantically unpredictable sentences (Benoit *et al.*, 1996) intended for use in intelligibility tests. The fairy tales were divided into sentences and read in the original order of the work. Each sentence was individually presented to the speaker using a flat panel monitor.

This corpus contains disjoint training and test sets. The total recording time for the training set is about 3.5 hours and it consists of about 3500 sentences: 1500 randomly chosen newspaper sentences, 1000 newspaper sentences chosen based on diphone coverage, and 1000 fairytale sentences. The recording time for the test set is about 0.5 hours and it comprises 200 randomly chosen newspaper sentences, 100 randomly chosen novel sentences and 200

[4] 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.

Table 1: Phonetic coverage of each subset of the RSS corpus.

Subset	Sentences	Size [min]	Diphones	Diphones/ sentence	Quinphones	Quinphones/ sentence
Random	1500	104	662	0.44	41285	27.5
Diphone	1000	53	706	0.71	26385	26.3
Fairytale	1000	67	646	0.65	29484	29.4

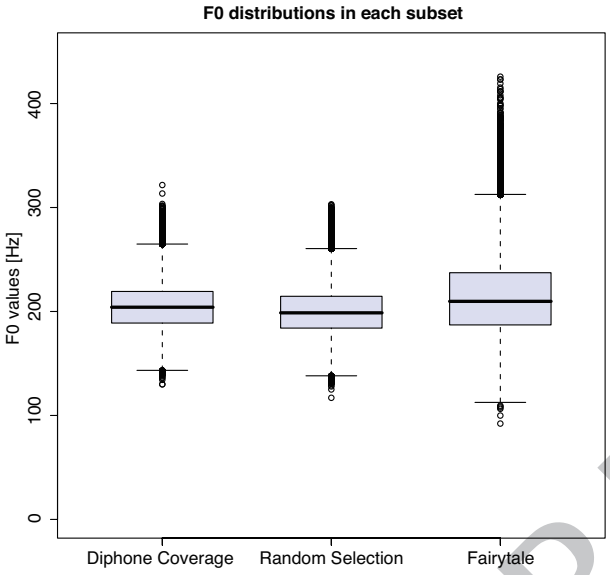


Figure 2: F_0 distributions in each subset.

semantically unpredictable sentences.

Table 1 shows the total number of different diphones and quinphones in these subsets. Diphones are the typical unit used for unit selection systems and quinphones are the base unit for HMM-based speech synthesis systems[5]. A larger number of types implies that the phonetic coverage is better. From the diphones/sentence column in the table we can see that the subset designed for diphone coverage has better coverage in terms of the number of different diphone types but – looking at the quinphones/sentence column – its coverage of quinphones is slightly worse than random selection . This indicates that the appropriate text design or sentence selection policy for HMM-based speech synthesis should be different from that for unit selection.

All recorded sentences were manually endpointed and have been checked for consistency against the orthographic form. The newspaper sentences were read out using a relatively flat intonation pattern, while the fairy tales had a more narrative rhythm and prosody. Figure 2 shows the box-plots of F_0 values extracted from all the sentences of each subset, in which the mean is represented by a solid bar across a box showing the quartiles, whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. From this figure we can see that the subset including fairy tales has wider F_0 variation than other subsets.

[5]The units are further extended by adding prosodic contexts mentioned in Section 3.

3. Romanian Front-end Text Processing

Text processing is one of the most challenging aspects of any new language for a text-to-speech system. The great variability among different language groups and local specific alterations to standard spelling or grammar make it an important and vital part of any TTS system.

For Romanian, there are a few projects and publications regarding text processing, such as (Burileanu *et al.*, 1999), (Frunza *et al.*, 2005). However, their availability and applicability is limited. For the purpose of this study, a new text processor was developed, based on the Cerevoice development framework (Aylett and Pidcock, 2007). Language-dependent data has been gathered and probabilistic models have been trained; the front-end outputs HTS format labels comprising 53 kinds of contexts (Zen *et al.*, 2007c). The following sections describe the resources used in developing the front-end.

3.1. Text corpus

We utilised newspaper articles obtained from the RSS feed of the Romanian free online newspaper, Adevarul. The articles were gathered over the period of August to September 2009 and they amount to about 4500 titles and over 1 million words. Due to the variety of character encodings used, the text corpus had to be cleaned and normalised before further processing.

3.2. Phonemes and letter-to-sound rules

The Romanian phonetic inventory generally consists of 7 vowels, 2 to 4 semivowels and 20 consonants. Table 2 shows the phone set used in our experiments. Romanian letter-to-sound rules are straightforward. However there are several exceptions, which occur mainly in vowel sequences, such as diphthongs and triphthongs. Therefore we adopted a lightly supervised automatic learning method for letter-to-sound rules as follows: From the text corpus, the top 65,000 most frequent words were extracted. General simple initial letter-to-sound rules were written manually by a native speaker. These rules were

Table 2: Phone set used in the experiments, given in SAMPA.

vowel	a @ 1 e i i_0 o u
semivowel	e_X j o w
nasal	m n
plosive	b d g k p t
affricate	ts tS dZ
fricative	f v s z S Z h
trill	r
approximant	l
silence/pause	*'sil' 'pau'*

used to phonetically transcribe the complete list of words. To deal with the exceptions above, the pronunciations of 1000 words chosen at random were checked, and corrected where necessary, by a native speaker. Using this partially-corrected dictionary of 65,000 words, letter-to-sound rules were automatically learned using a classification and regression tree (CART) (Breiman *et al.*, 1984). The accuracy of the obtained model is about 87%, measured using 5-fold cross validation. A small additional lexicon was manually prepared to deal mainly with neologisms, whose pronunciations are typically hard to predict from spelling.

3.3. Accent

Romanian has no predefined accentual rules. Different cultural and linguistic influences cause variation in the positioning of the accent across groups of related words. However, the online SQL database of the Romanian Explicative Dictionary (DEX: `http://dexonline.ro/`) provides accent positioning information. Using this information from DEX directly, an accent location dictionary for the 65,000 most frequent words in the text corpus was prepared.

3.4. Syllabification

Romanian syllabification has 7 basic rules, but these can be affected by morphology, such as compound words or hyphenated compounds. These rules apply to the orthographic form of the words. In our approach, we have used the maximal onset principle applied to the phonetic transcription of the words. Onset consonant groups and vowel nuclei have been defined. Based on partial evaluation of the principle, we determined that the accuracy of the syllabification is approximately 75% . One of the major exceptions occurs in the vowel-semivowel-vowel groups, where both the vowel-semivowel and semivowel-vowel group can be a diphtong, thus a nuclei. Another important exception is represented by the compound words, where the syllabification is based on morphological decomposition and not the standard rules.

3.5. Part-of-speech (POS) tagging

We used a Romanian POS tagger available online from `http://www.cs.ubbcluj.ro/~dtatar/nlp/WebTagger/WebTagger.htm`. Most of the text corpus was split into sentences and tagged using this tool. The accuracy of the POS tagging is 70% on average, according to internal evaluation results reported by the developers of the POS tagger.

3.6. HTS labels

HTS labels were generated using the text processor, based on the recorded sentences and scripts. All the words found in the recorded sentences were checked in the lexicon for correct phonetic transcription and accent location.

4. Building HMM-based speech synthesis systems using a high sampling frequency

We adopted a recent HMM-based speech synthesis system described in (Zen *et al.*, 2007a), which uses a set of speaker-dependent context-dependent multi-stream left-to-right state-tied (Young *et al.*, 1994; Shinoda and Watanabe, 2000) multi-space distribution (MSD) (Tokuda *et al.*, 2002) hidden semi-Markov models (HSMMs) (Zen *et al.*, 2007b) that model three kinds of parameters, required to drive the STRAIGHT (Kawahara *et al.*, 1999) mel-cepstral vocoder with mixed excitation (Kawahara *et al.*, 2001). Once we define context-dependent labels from the language-dependent front-end outputs, the framework of this system is basically language-independent and thus we can directly use it on our data.

The sampling frequency of the speech directly affects feature extraction and the vocoder and indirectly affects HMM training via the analysis order of spectral features. The following sections give an overview of how the sampling frequency affects the first-order all-pass filter used for mel-cepstral analysis and how we can utilise higher sampling frequencies in this analysis method.

4.1. The first-order all-pass frequency-warping function

In mel-cepstral analysis (Tokuda *et al.*, 1991), the vocal tract transfer function $H(z)$ is modelled by M-th order mel-cepstral coefficients $\boldsymbol{c} = [c(0), \ldots, c(M)]^\top$ as follows:

$$H(z) = \exp \boldsymbol{c}^\top \tilde{\boldsymbol{z}} = \exp \sum_{m=0}^{M} c(m) \tilde{z}^{-m}, \tag{1}$$

where $\tilde{\boldsymbol{z}} = [1, \tilde{z}^{-1}, \ldots, \tilde{z}^{-M}]^\top$. $\tilde{z}^{-1}$ is defined by a first-order all-pass (bilinear) function

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad |\alpha| < 1 \tag{2}$$

and the warped frequency scale $\beta(\omega)$ is given as its phase response:

$$\beta(\omega) = \tan^{-1} \frac{(1 - \alpha^2) \sin \omega}{(1 + \alpha^2) \cos \omega - 2\alpha}. \tag{3}$$

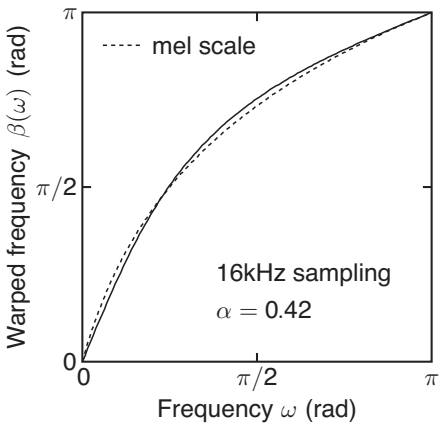


Figure 3: Frequency warping using the all-pass function. At a sampling frequency of 16 kHz, $\alpha = 0.42$ provides a good approximation to the mel scale.

The phase response $\beta(\omega)$ gives a good approximation to an auditory frequency scale with an appropriate choice of α.

An example of frequency warping is shown in Fig. 3. where it can be seen that, when the sampling frequency is 16 kHz, the phase response $\beta(\omega)$ provides a good approximation to the mel scale for $\alpha = 0.42$. The choice of α depends on the sampling frequency used and the auditory scale desired. The next section describes how to determine this parameter for a variety of auditory scales.

4.2. The Bark and ERB scales using the first-order all-pass function

In HMM-based speech synthesis, the mel scale is widely used. For instance, Tokuda *et al.* provide appropriate α values for the mel scale for speech sampling frequencies from 8kHz to 22.05kHz (Tokuda *et al.*, 1994b).

In addition to the mel scale, the Bark and equivalent rectangular bandwidth (ERB) scales (Patterson, 1982) are also well-known auditory scales. In (Smith III and Abel, 1999), Smith and Abel define the optimal α (in a least-squares sense) for each scale as follows:

$$\alpha_{\mathrm{Bark}} = 0.8517\sqrt{\arctan(0.06583\, f_s)} - 0.1916 \quad (4)$$

$$\alpha_{\mathrm{ERB}} = 0.5941\sqrt{\arctan(0.1418\, f_s)} + 0.03237 \quad (5)$$

where f_s is the waveform sampling frequency. However, note that the error between the true ERB scale and all-pass scale approximated by α_{ERB} is three times larger than the error for the Bark scale using α_{Bark} (Smith III and Abel, 1999). Note also that as sampling rates become higher, the accuracy of approximation using the all-pass filter becomes worse for both scales.

4.3. HMM training

The feature vector for the MSD-HSMMs consists of three kinds of parameters: the mel-cepstrum, generalised

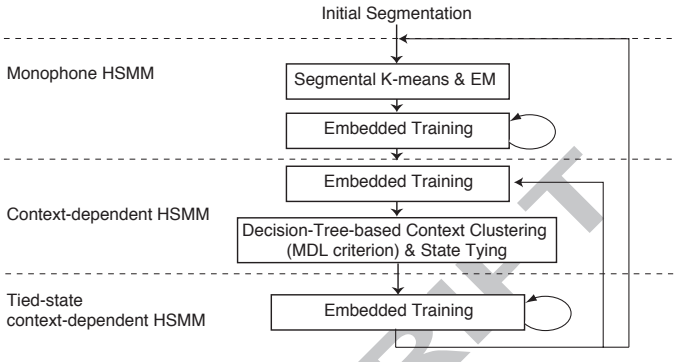


Figure 4: Overview of HMM training stages for HTS voice building.

$\log F_0$ (Yamagishi and King, 2010) and a set of band-limited aperiodicity measures (Ohtani *et al.*, 2006), plus their velocity and acceleration features.

An overview of the training stages of the HSMMs is shown in Figure 4. First, monophone MSD-HSMMs are trained from the initial segmentation using the segmental K-means and EM algorithms (Dempster *et al.*, 1977), converted to context-dependent MSD-HSMMs and re-estimated using embedded training. Then, decision-tree-based context clustering (Young *et al.*, 1994; Shinoda and Watanabe, 2000) is applied to the HSMMs and the model parameters of the HSMMs are thus tied. The clustered HSMMs are re-estimated again using embedded training. The clustering processes are repeated until convergence of likelihood improvements (inner loop of Figure 4) and the whole process is further repeated using segmentation labels refined with the trained models in a bootstrap fashion (outer loop of Figure 4). In general, speech data sampled at higher rates requires a higher analysis order for mel-cepstral analysis. We therefore started by training models on lower sampling rate speech (16 kHz) with a low analysis order and gradually increased the analysis order and sampling rates via either re-segmentation of data or single-pass retraining of HMMs (Yamagishi and King, 2010).

4.4. Configurable parameters

In order to establish a benchmark system which will be useful for many future experiments, we carefully adjusted various configurable parameters as follows:

1. From initial analysis-by-synthesis tests using five sentences followed by informal listening, we first chose the spectral analysis method and order. Specifically, we compared mel-cepstrum and mel-generalised cepstrum (MGC) (Tokuda *et al.*, 1994a) at orders of 50, 55, 60, 65 and 70, using Bark and ERB frequency warping scales[6] using speech data sampled at 48 kHz. The parameter to control all-pole or cepstral analysis

[6]Strictly speaking, we should call them Bark-cepstrum and ERB-cepstrum. However, for simplicity we will just call them all 'mel-cepstrum'.

method was set to 3 (Tokuda *et al.*, 1994a). The results indicated the use of MGC with 60th order and the Bark scale. However, the differences between the Bark and ERB scales were found to be not as great as differences due to the sampling frequency. Our earlier research (Yamagishi and King, 2010) also found that the auditory scale – including the Mel scale – was not a significant factor. Therefore we omitted the ERB scale and the Mel scale from the listening test reported later. We repeated the same process for speech data sampled at 32 kHz and chose MGC with 44th order with the Bark scale.

2. Preliminary HMM training was then carried out to determine training data partitions. A total of 20 systems resulted from combinations of the recorded data used in sets of 500, 1000, 1500, 2500 and 3500 sentences. From informal listening, the fairy tale sentences were found to alter the overall quality of the synthesised speech, since these sentences had a more dynamic prosody than the newspaper sentences (see Figure 2). Therefore we excluded the fairy tale set and used a 2500 sentence set in subsequent experiments.
3. We employed the data-driven generalised-logarithmic F_0 scale transform method proposed in (Yamagishi and King, 2010). The maximum likelihood estimator for the generalised logarithmic transform obtained from F0 values of all voiced frames included in the RSS database, using the optimisation method mentioned in (Yamagishi and King, 2010), was 0.333.
4. We then separated decision trees for speech from non-speech units (pauses and silences) rather than having a shared single tree.

In the experiments reported in this paper, only speech recorded using the Sennheiser MKH 800 microphone was used. Investigation of the differences caused by microphone type are left as future work.

5. Evaluation

5.1. Listening Test

For the listening test, we used the framework from the Blizzard Challenge (Karaiskos *et al.*, 2008) and evaluated speaker similarity, naturalness and intelligibility.

We recruited a total of 54 Romanian native listeners of which 20 completed the test in purpose-built, soundproof listening booths and the rest evaluated the systems on their personal computers and audio devices, mostly using headphones. They each evaluated a total of 108 sentences randomly chosen from the test set, 36 from each category (news, novel, SUS). The speaker similarity and naturalness sections contained 18 newspaper sentences and 18 novel sentences each. 36 SUSs were used to test intelligibility.

The duration of the listening test was about 45 minutes per listener. Listeners were able to pause the evaluation at any point and continue at a later time, but the majority opted for a single listening session. Most of the listeners had rarely listened to synthetic voices; they found the judgement of naturalness and speaker similarity to be the most challenging aspects of the test.

Nine individual systems were built for the evaluation. All used the same front-end text processing. They differ in the synthesis method used (HMM-based, unit selection), sampling frequency (16 kHz, 32 kHz, 48 kHz) and the amount of data used for the training of the voice. The analysis of the three microphones is an interesting topic but, in order to make the listening tests feasible, we had to exclude this factor. The systems are identified by letter:

A Original recordings, natural speech at 48 kHz

B Unit selection system at 16 kHz, using 3500 sentences

C Unit selection system at 32 kHz, using 3500 sentences

D Unit selection system at 48 kHz, using 3500 sentences

E HMM system at 48 kHz, using 500 training sentences

F HMM system at 48 kHz, using 1500 training sentences

G HMM system at 16 kHz, using 2500 training sentences

H HMM system at 32 kHz, using 2500 training sentences

I HMM system at 48 kHz, using 2500 training sentences

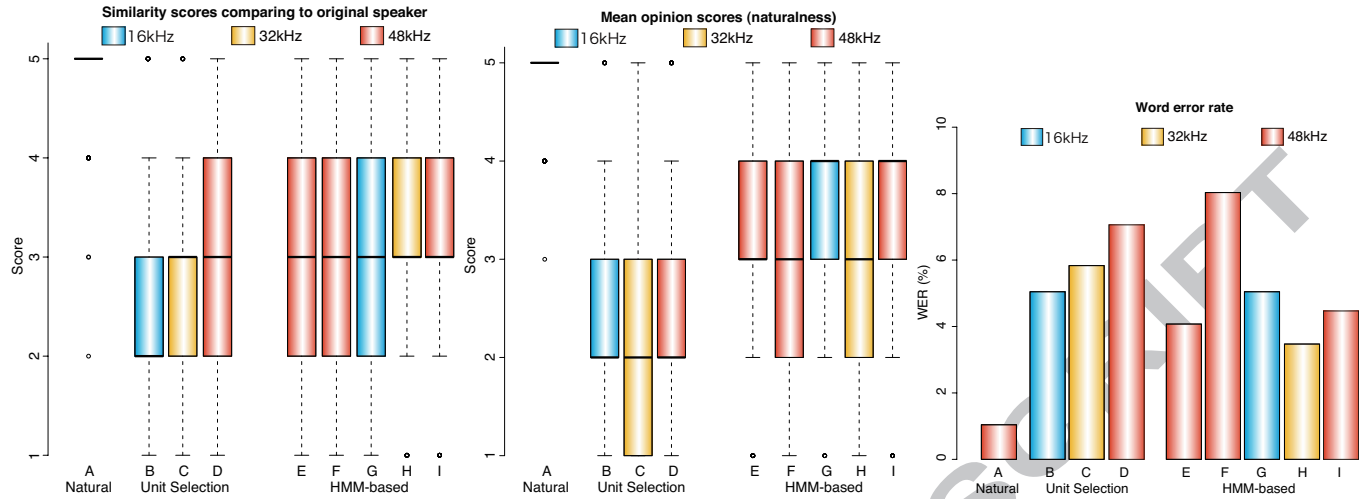
By comparing systems B, C and D with E, F, G, H and I, we can see the effect of the synthesis method. By comparing systems B,C,D or G,H,I, we can see the effect of sampling frequency, per synthesis method. Comparing systems E,F,I, we can see the effect of the amount of training data for the HMMs.

In the speaker similarity task, after the listeners listened to up to 4 original recording samples, they were presented with a synthetic speech sample generated from one of the nine systems and were asked to rate similarity to the original speaker using a 5-point scale. The scale runs from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person]. In the naturalness evaluation task, listeners used a 5-point scale from 1 [Completely Unnatural] to 5 [Completely Natural]. In the intelligibility task, the listeners heard a SUS and were asked to type in what they heard. Typographical errors and spelling mistakes were allowed for in the scoring procedure. The SUS each comprised a maximum of 6 frequently-used Romanian words.

5.2. Results

5.2.1. Speaker similarity

The left column of Fig. 5 shows the results for speaker similarity. We first observe a clear separation between the original voice (system A), HMM voices (systems E, F, G, H and I) and unit selection voices (systems B, C and D). We can also observe a clear influence of the sampling frequency over speaker similarity although improvements seem to level off at 32kHz. This is a new and interesting finding. Also there is some influence of the amount of training data. We can see that the difference between



Mean scores

A	B	C	D	E	F	G	H	I
4.9	2.6	2.5	2.7	3.1	3.0	3.1	3.4	3.3

Mean scores

A	B	C	D	E	F	G	H	I
4.8	2.2	2.5	2.4	3.3	3.0	3.4	3.4	3.4

Mean Score [**WER**%]

A	B	C	D	E	F	G	H	I
1.0	5.0	5.8	7.1	4.1	8.0	5.0	3.5	4.5

Significance at 1% level for **similarity**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	1	1	1	1	1
C	1	0	-	0	1	1	1	1	1
D	1	0	0	-	1	0	0	1	1
E	1	1	1	1	-	0	0	0	0
F	1	1	1	0	0	-	0	1	1
G	1	1	1	0	0	0	-	1	0
H	1	1	1	1	0	1	1	-	0
I	1	1	1	1	0	1	0	0	-

Significance at 1% level for **naturalness**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	1	1	1	1	1
C	1	0	-	0	1	1	1	1	1
D	1	0	0	-	1	1	1	1	1
E	1	1	1	1	-	1	0	0	0
F	1	1	1	1	1	-	1	1	1
G	1	1	1	1	0	1	-	0	0
H	1	1	1	1	0	1	0	-	0
I	1	1	1	1	0	1	0	0	-

Significance at 1% level for **WER**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	0	1	1	0	1
B	1	-	0	0	0	0	0	0	0
C	1	0	-	0	0	0	0	0	0
D	1	0	0	-	0	0	0	0	0
E	0	0	0	0	-	0	0	0	0
F	1	0	0	0	0	-	0	0	0
G	1	0	0	0	0	0	-	0	0
H	0	0	0	0	0	0	0	-	0
I	1	0	0	0	0	0	0	0	-

Figure 5: Listening tests results. There are three columns of plots and tables which are, from left to right, similarity to original speaker, mean opinion score for naturalness, and intelligibility. The similarity and naturalness plots on the upper row are box plots where the median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range. The three tables in the middle row give the mean scores of each system. The tables in the bottom row indicate significant differences between pairs of systems, based on Wilcoxon signed rank tests with alpha Bonferoni correction (1% level); '1' indicates a significant difference.

systems E and F is less significant whereas the difference between systems F and I is significant. We believe that neither 500 nor 1500 sentences were sufficient for training models that can reproduce good speaker similarity, since our feature dimension is very high due to the high order mel-cepstral analysis.

Although we expected that unit selection would have better similarity than HMM-based, the results are contrary to our expectation. This may be explained by the corpus design: In our corpus, only 1000 sentences were chosen based on diphone coverage and the remaining 2500 sentences consist of 1500 randomly chosen newspaper sentences and 1000 fairy tale sentences. Even if we combine both types of sentence, there are still 16 missing diphones and 79 diphones having fewer than 3 occurrences. Although quinphones, the base unit of HMM voices, do not have good coverage either, unit selection systems (which use diphone units) are known to be more sensitive to lack of phonetic coverage, compared to HMM-based systems (Yamagishi *et al.*, 2008b).

5.2.2. Naturalness

We can see similar tendencies to those for the similarity task, except that sampling frequency does not seem to have any effect. The use of higher sampling frequency did not improve the naturalness of synthetic speech, in contrast to speaker similarity. This is also an interesting finding. Regarding the amount of data, we see that there are some fluctuations, although the largest amount of data typically leads to the best voice for each synthesis method.

5.2.3. Intelligibility

Unfortunately there appears to be something of a ceiling effect on intelligibility. Absolute values of WER are generally small: both synthesis methods have good intelligibility. Even though we observe that systems D and

F have a slightly higher error rate, there are no statistically significant differences between any pairs of synthetic voices in terms of WER. To confirm this we performed a small additional test including paronyms and obtained the same results. We believe that the lack of significant differences between systems is partly caused by the nature of the simple grapheme-to-phoneme rules in Romanian. Even for SUSs and paronyms, both natural and synthetic speech are easy to transcribe, leading to WERs close to zero. This result suggests there is a need for better evaluation methods for the intelligibility of synthetic speech in languages such as Romanian.

5.2.4. Listening environments

We performed an ANOVA test to discover whether the listening environment affects the results. An ANOVA test at 1% significance level shows that only the system C (unit selection system at 32 kHz, using 3500 sentences) in the similarity test was affected by the listening environment. The subjects who completed the test in the listening booths generally gave lower similarity scores for system C.

5.2.5. Summary

This RSS corpus is probably better suited to HMM-based synthesis than to unit selection. All speech synthesis systems built using the corpus have good intelligibility. However, we need to design a better evaluation of the system's intelligibility in simple grapheme-to-phoneme languages such as Romanian.

We found that the sampling frequency is an important factor for speaker similarity. More specifically, downsampling speech data in this corpus to 32kHz does no harm, but downsampling to 16kHz degrades speaker similarity. The use of higher sampling frequency, however, did not improve either the naturalness or intelligibility of synthetic speech.

These results are consistent with existing findings: (Fant, 2005) mentions that almost all the linguistic information from speech is in the frequency range 0 to 8 kHz. This implies that a 16 kHz sampling frequency (and thus 8 kHz Nyquist frequency) is sufficient to convey the linguistic information. Our results also shown that using sampling frequencies over 16 kHz did not improve the intelligibility of synthetic speech. On the other hand, a classic paper regarding sampling frequency standardisation (Muraoka *et al.*, 1978) reported that a cut-off frequency of less than 15 kHz may deteriorate audio quality. This means that the sampling frequency used should be higher than 30 kHz. In fact, our results do show that downsampling to 16kHz degrades speaker similarity. Therefore we can conclude that the naturalness and intelligibility of synthetic speech only require transmission of linguistic information, which can be achieved at 16kHz sampling frequency, whereas speaker similarity of synthetic speech is affected by audio quality (requiring a higher sampling rate).

5.3. Demos

We encourage interested readers to listen to audio samples comprising some of the materials used for listening tests `http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/rss.html` and the first 3 chapters of a Romanian public-domain novel "Moara cu noroc" by Ioan Slavici, available online via `http://octopus.utcluj.ro:56337/moaraCuNoroc/moaraCuNoroc.rss`. We also encourage them to test our live demo `http://octopus.utcluj.ro:56337/HTS_RomanianDemo/index.php`. The RSS database itself can be downloaded from `http://octopus.utcluj.ro:56337/RORelease/`.

6. Conclusions

This paper has introduced a newly-recorded high-quality Romanian speech which we call "RSS", along with Romanian front-end modules and HMM-based voices. In order to promote Romanian speech technology research, all of these resources are freely available for academic use.

From the listening tests presented here, we conclude that 1) the RSS corpus is well-suited for HMM-based speech synthesis and 2) that the speech synthesis systems built from the corpus have good intelligibility.

Using the RSS corpus, we have also revisited some basic configuration choices made in HMM-based speech synthesis such as the sampling frequency and auditory scale, which have been typically chosen based on experience from other fields. We found that higher sampling frequencies (above 16 kHz) improved speaker similarity. More specifically, the speech data in this corpus can be downsampled to 32kHz without affecting results but that downsampling to 16 kHz degrades speaker similarity.

Future work includes an analysis of each of the three microphones used and designing a better intelligibility evaluation for the simple grapheme-to-phoneme languages, such as Romanian.

7. Acknowledgements

A simplified description of some of this research was published in (Yamagishi and King, 2010).

Adriana Stan is funded by the European Social Fund, project POSDRU/6/1.5/S/5 and was visiting CSTR at the time of this work. Junichi Yamagishi and Simon King are partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF – http://www.ecdf.ed.ac.uk). The ECDF is partially supported by the eDIKT initiative (http://www.edikt.org.uk).

We would like to thank CereProc staff for support with the text processing tools, Catalin Francu for assistance

with the DEX-online database, and the authors of the Romanian POS Tagger. The first author would also like to thank everyone at CSTR – especially Oliver Watts – for their support and guidance.

References

Aylett, M. and Pidcock, C. (2007). The CereVoice characterful speech synthesiser SDK. In *Proc. AISB 2007*, pages 174–178, Newcastle, U.K.

Benoit, C., Grice, M., and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, **18**(4), 381–392.

Black, A. and Cambpbell, N. (1995). Optimising selection of units from speech database for concatenative synthesis. In *Proc. EUROSPEECH-95*, pages 581–584.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Burileanu, D., Dan, C., Sima, M., and Burileanu, C. (1999). A parser-based text preprocessor for Romanian language TTS synthesis. In *Proc. EUROSPEECH-99*, pages 2063–2066, Budapest, Hungary.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1), 1–38.

Fant, G. (2005). *Speech Acoustics and Phonetics: Selected Writings*, chapter Speech Perception, pages 199–220. Springer Netherlands.

Ferencz, A. (1997). *Contributii la dezvoltarea sintezei text-vorbire pentru limba romana*. Ph.D. thesis, University of Cluj-Napoca.

Frunza, O., Inkpen, D., and Nadeau, D. (2005). A text processing tool for the romanian language. In *Proc. EuroLAN 2005: Workshop on Cross-Language Knowledge Induction*, Cluj-Napoca.

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP-96*, pages 373–376.

Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C. (2008). The Blizzard Challenge 2008. In *Proc. Blizzard Challenge Workshop*, Brisbane, Australia.

Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, **27**, 187–207.

Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *2nd MAVEBA*.

Muraoka, T., Yamada, Y., and Yamazaki, M. (1978). Sampling-frequency considerations in digital audio. *J. Audio Eng. Soc*, **26**(4), 252–256.

Ohtani, Y., Toda, T., Saruwatari, H., and Shikano, K. (2006). Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In *Proc. Interspeech 2006*, pages 2266–2269.

Patterson, R. (1982). Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, **76**, 640–654.

Shinoda, K. and Watanabe, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *J. Acoust. Soc. Japan (E)*, **21**, 79–86.

Smith III, J. O. and Abel, J. S. (1999). Bark and ERB bilinear transforms. *IEEE Trans. on Speech Audio Process.*, **7**(6), 697–708.

Tokuda, K., Kobayashi, T., Fukada, T., Saito, H., and Imai, S. (1991). Spectral estimation of speech based on mel-cepstral representation. *IEICE Trans. Fundamentals*, **J74-A**(8), 1240–1248. in Japanese.

Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. (1994a). Mel-generalized cepstral analysis — a unified approach to speech spectral estimation. In *Proc. ICSLP-94*, pages 1043–1046, Yokohama, Japan.

Tokuda, K., Kobayashi, T., and Imai, S. (1994b). Recursive calculation of mel-cepstrum from LP coefficients. In *Technical Report of Nagoya Institute of Technology*.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (2002). Multi-space probability distribution HMM. *IEICE Trans. Inf. & Syst.*, **E85-D**(3), 455–464.

Yamagishi, J. and King, S. (2010). Simple methods for improving speaker-similarity of HMM-based speech synthesis. In *Proc. ICASSP 2010*, pages 4610–4613, Dallas, TX.

Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., and Tokuda, K. (2008a). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In *Proc. Blizzard Challenge 2008*, Brisbane, Australia.

Yamagishi, J., Ling, Z., and King, S. (2008b). Robustness of HMM-based speech synthesis. In *Proc. Interspeech 2008*, pages 581–584, Brisbane, Australia.

Young, S., Odell, J., and Woodland, P. (1994). Tree-based state tying for high accuracy acoustic modeling. In *Proc. ARPA Human Language Technology Workshop*, pages 307–312.

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007a). Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. & Syst.*, **E90-D**(1), 325–333.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2007b). A hidden semi-Markov model-based speech synthesis system. *IEICE Trans. Inf. & Syst.*, **E90-D**(5), 825–834.

Zen, H., Nose, T., Yamagishi, J., Sako, S., and Tokuda, K. (2007c). The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pages 294–299.

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, **51**(11), 1039–1064.

Zwicker, E. and Scharf, B. (1965). A model of loudness summation. *Psych. Rev.*, **72**, 2–26.