TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision

A. Stan[1], *O. Watts*[2], *Y. Mamiya*[2], *M. Giurgiu*[1], *R. A. J. Clark*[2], *J. Yamagishi*[2,3], *S. King*[2]

[1]Communications Department, Technical University of Cluj-Napoca, Romania
[2]The Centre for Speech Technology Research, University of Edinburgh, UK
[3]National Institute of Informatics, Japan

{adriana.stan, mircea.giurgiu}@com.utcluj.ro

{owatts, Yoshitaka.Mamiya, robert, jyamagis, Simon.King}@inf.ed.ac.uk

Abstract

Simple4All Tundra (version 1.0) is the first release of a standardised multilingual corpus designed for text-to-speech research with imperfect or found data. The corpus consists of approximately 60 hours of speech data from audiobooks in 14 languages, as well as utterance-level alignments obtained with a lightly-supervised process. Future versions of the corpus will include finer-grained alignment and prosodic annotation, all of which will be made freely available. This paper gives a general outline of the data collected so far, as well as a detailed description of how this has been done, emphasizing the minimal language-specific knowledge and manual intervention used to compile the corpus. To demonstrate its potential use, text-to-speech systems have been built for all languages using unsupervised or lightly supervised methods, also briefly presented in the paper.

Index Terms: multilingual corpus, light supervision, imperfect data, found data, text-to-speech, audiobook data

1. Introduction

Building a text-to-speech (TTS) conversion system for a new language has in the past been an expensive and time-consuming activity. Using data-driven methods to build, for example, a statistical parametric waveform generation module or TTS *back-end*, can alleviate to some extent the lack of expert linguistic knowledge. Even then, however, a recording script must be prepared, a voice talent recruited and high-quality speech recording carefully supervised. Also problematic is the text-processing component of the system, i.e. the TTS *front-end*, if none is available for the target language. A front-end is made up of rule-based or statistical modules; acquiring the expert knowledge required either to manually specify those rules, or to annotate a learning sample on which to train the statistical models, represents a major obstacle to creating a TTS system for a new target language and requires highly specialised knowledge. Such non-trivial tasks include, for example, specifying a phoneme-set or part of speech (POS) tag-set for a language where one has not already been defined; annotating plain text with POS tags, as required to train a POS tagger and annotating the surface forms of words with phonemes to build a pronunciation lexicon.

One of the primary goals of the project *Simple4All*[1] is to produce freely available tools for building TTS systems with little or no expert supervision from freely available existing data. These tools enable us to sidestep the expense associated with engineering a speech corpus in each new target language from scratch, in the case where data is not readily available. Our toolkit includes modules for handling imperfect recording conditions, segmenting audio into manageable chunks, and aligning those chunks with a chapter- or book-level text transcription. We here explain how these tools have been applied to existing audiobook data in 14 languages, most of it freely available, to create a multilingual corpus with minimal manual intervention and language-specific expert knowledge.

The result of this processing is a standardised multilingual database of 'found' data, which we release under the name Tundra. There has been much recent interest in in using found data to produce TTS systems, in particular, speech data from audiobook recordings [1, 2, 3, 4, 5, 6, 7]. We note that the Arctic databases [8] have provided a valuable resource for research into TTS using conventional purpose-recorded databases, in that they are freely available and serve as a common point of reference for benchmarking. In view of this significant and growing interest in building TTS systems from found data, we feel there is a need for a similarly standardised and freely-available corpus of found data. We present Tundra to the TTS research community in the hope that it can start to fill that need.

Our toolkit also includes modules for selecting a subset of utterances with a uniform speaking style, and constructing TTS systems from text and speech data without reliance on language-specific expert knowledge or on conventional linguistic resources such as lexicons, phonesets, part-of-speech taggers etc. In order to show that it is feasible to build voices on corpora built with such minimal expert supervision, we also present a demonstration of TTS systems that we have built by applying these tools to Tundra. We do not present detailed explanation, evaluation and analysis of these demo systems here due to space limitations, and refer interested readers to [9], where such details will be given.

An initial public version of the Simple4All tools used to compile the corpus and build the demo voices is due to be released in November 2013.

2. Corpus Construction

In this section we describe the pipeline of data processing involved in building the Tundra corpus, from speech denoising and deverberation to lightly supervised speech and text alignment. All the steps presented in the following subsections are based solely on found speech and text resources and could be easily applied to any other resource, even by non-expert users. As regards language dependency, the only step which requires

[1]www.simple4all.org/

familiarity with at least the script of the target language is the first step of matching 10 minutes of speech with an orthographic transcript. All the other processes can be performed by the users with little or no training in speech processing and without relying on any target language knowledge.

2.1. Speech Pre-processing

Conventional TTS corpora deliver speech recorded in noise-free non-reverberant environments, and thus lead to high-quality synthetic speech. Found data, on the other hand are usually recorded in sub-optimal conditions, and without professional recording equipment. Therefore, when building TTS systems on this type of data, some pre-processing steps are in order.

For Tundra, recordings which casual listening suggested were sub-optimal went through the following pre-processing steps, applied to each recording session individually,[2] so that variations in between them can be normalised: 1) **Noise reduction** - uses a multi-band noise gate removal with a 20dB noise reduction threshold, a frequency smoothing of 150 Hz and 0.15 second decay time. The noise profile was selected from the initial silence segments of each speech file. 2) **Normalisation** - DC offset was removed, and the recordings were normalised to a maximum amplitude of -0.1 dB, so that the average energy level is the same across different recording sessions. 3) **Deverberation** - was performed using a RMS based algorithm, with a smoothing of 40 ms and a release of 400ms.

2.2. Lightly-supervised Audio Segmentation

Current parametric TTS systems generally use training data which is segmented into sentence-length chunks, and rarely make use of contexts beyond the current sentence. The small length of the training data is also a limitation of the forced alignment algorithm while training. Although several algorithms [4, 10, 11] have been proposed to enable the use of longer speech segments, we still consider that sentence-length utterances are the building blocks of TTS, and longer segments can be easily obtained by concatenating the former, thus ensuring a paragraph or maybe chapter level analysis or training.

[12] presents a lightly supervised method for the segmentation of speech into sentences. The method uses a small amount of manually labelled data, in which the silence between sentences is marked for around 5 to 10 minutes of speech. Silence marking is a trivial task and requires no technical knowledge.

Using the initial training data, standard Gaussian mixture models (GMMs) with 16 components are trained for speech and silence respectively. The observation vectors consist of energy, 12 dimensional MFCCs, their delta features, and the number of zero crossings in a frame. The distinction between speech and silence is made by calculating the log likelihood ratio (LLR) of each frame. The framewise LLR is smoothed using a moving median filter.

While doing sentence level segmentation, an important aspect is to discriminate between within-sentence breaks, and sentence boundary breaks. Therefore, the trained GMMs likelihood scores are evaluated on the training data, and the durations of the sentence boundary silence segments and the durations of within-sentence silence segments are computed. Two Gaussian PDFs are then fitted to the two model durations. The intersection point of the two PDFs is used as a duration threshold to classify silent segments as either sentence-internal or sentence boundary breaks.

Results presented in [12] showed that this method when applied to an English audiobook, successfully identified most of the sentence boundaries. We also evaluate it in this paper by comparing speech-based segmentation results against the text based ones.

2.3. Lightly-supervised Speech and Text Alignment

In [13] we first introduced a method for the automatic alignment of speech data with unsynchronised, imperfect transcripts, for a domain where no initial acoustic models are available. As opposed to [7], where existing high-quality acoustic and language models are used, our method requires only relatively low-quality grapheme-based acoustic models trained solely on the speech resource to be aligned. To overcome the lack of good acoustic models, the ASR decoding network is limited to a sequence of words derived from the approximate transcript, similar to [14]. This sequence is called a skip network. The confidence of the alignment is ranked based on the acoustic scores obtained in the decoding process with different degrees of freedom included in the skip network.

Manual intervention is limited to matching the first 10 minutes of speech with the correct text transcription, to provide data for training the initial acoustic models, similar to [15]. This feature makes the method easily applicable in any language employing an alphabetic writing system, and enables the use of found data without the hassle of manually transcribing its entirety.

Initial results on the English audiobook *A Tramp Abroad* by Mark Twain[3] showed an average 55% confident data, with a WER of 1% and SER of 8%. Since then, the acoustic model training has been extended to tri-grapheme and lightly supervised discriminative training [16], which led to an average of 75% confident data with similar word and sentence error rates. One major loss in sentence accuracy rates is due to utterance initial and final word deletions and insertions, which cannot be correctly detected by the current confidence measure. However, previous studies [17] showed that phone errors less than 1% do not degrade the quality of the synthetic speech.

The output of the alignment process is a set of segmented speech files with their corresponding orthographic transcripts, including punctuation, and also a time alignment of the segments within the initial speech data.

3. The Corpus

The procedures described above have been applied to a number of freely available found resources. Audiobooks were a first choice, as they are a readily available in multiple languages and are generally read by a single speaker and recorded with equipment of at least reasonable quality. Another advantage would be that by using cohesive and expressive spoken data as the basis for training a TTS system might yield more cohesive and expressive multi-utterance TTS output, fact which explains the high interest in them lately. This latter advantage is not especially made use of in the demo voices presented here, but is the subject of on-going work for us elsewhere.

To emphasise the utility of audiobooks in TTS systems, in Fig. 1 we present a comparison between standard TTS corpora and audiobooks with respect to logF0 in 4 different languages. The standard TTS corpora are: a subset of the database called 'Nina' in [18], a subset of a corpus of Finnish speech recorded

[2] Audiobooks are usually distributed in chapter-size chunks which correspond to one recording session.

[3] http://librivox.org/a-tramp-abroad-by-mark-twain/

Table 1: Simple4All Tundra Corpus overview

Language	Code	Author	Title	Speaker gender	Total [hours]	SNR [dB]	#Utts VAD	#Utts text	Aligned [hours]	Percent [%]
Bulgarian	BG	Yordan Yovkov	Zhetvariat	M	6.1	65	3139	4379	4.1	67.21
Danish	DA	J. & W. Grimm	Grimms eventyr I udvalg	M	2.1	33	1099	1112	1.1	52.38
Dutch	NL	Leo Tolstoy	Anna Karenina	M	6.5	42	3844	2405	4.9	75.38
English	EN	Stella Benson	Living Alone	F	4.5	64	2194	2632	2.4	53.33
Finnish	FN	Juhani Aho	Rautatie	F	3.1	40	1357	1673	2.6	83.87
French	FR	Voltaire	Candide ou L'optimisme	M	4	45	1890	1661	2.3	57.50
German	DE	Oscar Wilde	Das Bildnis des Dorian Gray	M	9.5	40	4865	4623	8	84.21
Hungarian	HU	Geza Gardonyi	Egri csillagok	F	8.5	38	4510	8375	5	58.82
Italian	IT	Anton Giulio Barrili	Galatea	M	6.5	55	2241	3874	5	76.92
Polish	PL	Wladyslaw Orkan	Siedem wybranych opowiadan	F	3.1	39	2078	2027	2.9	93.55
Portuguese	PR	Jose de Alencar	Senhora	F	9.3	29	5001	4740	5.2	55.91
Romanian	RO	Ioan Slavici	Mara	F	11.1	56	5563	6285	7	63.06
Russian	RU	Leo Tolstoy	Ucheniye Khrista	M	2.1	52	1113	1426	1.6	76.19
Spanish	ES	Miguel de Cervantes	Don Quijote de la Mancha	M	12.1	54	7902	5569	8	66.11

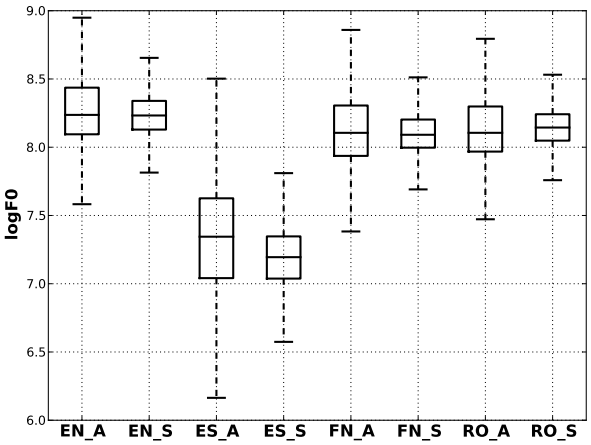

Figure 1: logF0 comparison of conventional TTS corpora versus audiobook data in four languages: English (EN), Spanish (ES), Finnish (FN) and Romanian (RO). **A** denotes the audiobook data, and **S** denotes the standard TTS database. The standard corpora speaker genders are the same as the selected audiobooks.

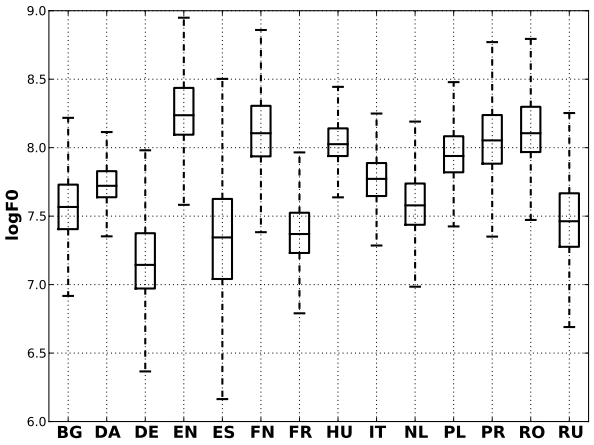

Figure 2: logF0 boxplots for all languages. Language codes are given in Table 1

from a female speaker specifically for TTS purposes, SEV neutral [19] and RSS [20]. It can be easily observed that the audiobooks have a greater standard deviation compared with conventional corpora, which means that they could easily provide a much richer prosodic context. This aspect can also be noticed from Fig. 2 where logF0 distributions are plotted for all the languages of the corpus.

As a result, Tundra 1.0 includes 14 audiobooks in 14 languages: Bulgarian, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Polish, Portuguese, Romanian, Russian and Spanish. Language selection was based on the availability of both speech and text data, as well as the language having an alphabetic writing system (in this case, Latin and Cyrillic alphabets). Important resources for these are the Librivox and Gutenberg[4] projects, which are the sources for most of the data used to compile Tundra. The complete list speech and text sources can be found here `http://tundra.simple4all.org/`. Table 1 presents an overview of the entire corpus, including title and author of the audiobook, speaker gender and total duration. There are 8 male and 6 female speakers, and the aligned corpus amounts to approximately 60 hours of speech.

For the final set of utterances included in this corpus, each audiobook underwent the steps described in the Section 2 and which are schematically depicted in Fig. 3. Audiobook chapters were converted from mp3 to *wav* format and then cleaned if the overall quality was considered low.[5] The first 10 minutes of speech were then annotated with silence segments and manually transcribed. Manual transcription proved to be a trivial task, and based on the book text, the authors were able to perform it, although they do not speak most of the languages included in the corpus. For the Cyrillic writing system languages (i.e. Bulgarian and Russian), native speakers were asked to correct an initial transcription provided by the authors. Data was then segmented using the VAD algorithm, and the resulting number of speech utterances is presented in Table 1 alongside the text-based segmentation. The difference between the number of VAD and text

[4] `http://librivox.org` and `http://gutenberg.org/`

[5] For example, the Spanish and Romanian data are professional recordings which did not require any pre-processing.

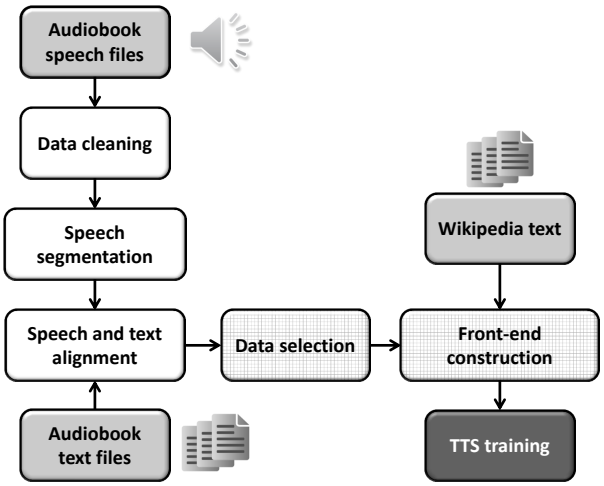

Figure 3: Outline of corpus construction and voice building

utterances results from the writing style of the book (i.e. mostly dialogue, or mostly descriptive) and the fact that in the alignment process, in order to obtain the most data from the audiobook, segmented utterances which are shorter than a specified threshold (5 seconds for these data) are concatenated.

After the alignment process, an average of 68% of the data were considered confident and included in the final corpus. Table 1 presents the duration of the aligned data and its percentage from the total duration. This percentage appears to be highly dependent on: a) the total amount of data available: see the low percentage of the Danish audiobook which has only 2.1 hours; b) speaker gender: female voices seem to have a lower alignment percentage; c) grapheme-to-phoneme language complexity: see English and French versus Italian and German;[6] and d) speaker characteristics: speaking rhythm, degree of expresivity, as well as general voice quality also affect the results.[7]

SER and WER values for the aligned audiobooks could not be exactly determined, as this would have required their full manual transcription, which is outside the scope of this corpus building procedure. However, one chapter from each audiobook in the languages spoken by the authors was evaluated, and the errors tend to be similar to those in [13], meaning a less than 1% WER and a 8% SER. Higher error rates were reported for the noisier speech data (see Table 1 for general signal-to-noise ratios).

To be useful as a standardised TTS corpus, Tundra is also partitioned into training and test sets. To ensure a satisfactory amount of testing data even for the shortest audiobook, the test data were selected from the final chapters/parts of the audiobooks, so that they amount to at least 10% of the aligned duration of it. The entire segmented and aligned corpus, along with the chapter-wise time alignment and training/test set division of can be downloaded from `http://tundra.simple4all.org`.

[6]Spanish and Romanian also have very simple G2P rules, but the speakers' greater expressivity limits the alignner's performance.

[7]This being a subjective measure, we encourage readers to listen to samples of the audiobooks.

4. Demo

To show the feasibility of using a corpus that has been compiled with such minimal intervention and language-specific expertise, we have used it to build demo TTS voices in the corpus languages. To build these voices we first select a subset of utterances spoken in a homogenous style using a slightly supervised active learning-based approach. We then employ a toolkit which has been specifically designed to construct TTS front-ends while making as few implicit assumptions about the target language as possible, and to be configurable with minimal effort and expert knowledge to suit arbitrary new target languages. The modules of our toolkit therefore rely where possible on resources which are intended to be universal. For example, to tokenise input text we rely on character properties given in the Unicode character database – a regular expression defined over these properties has so far produced sensible tokenisations in a variety of alphabetic (Latin-based, Cyrillic) and alphasyllabic (Brahmic) scripts.

A letter-based approach is used, in which the names of letters are used directly as the names of speech modelling units (in place of the phonemes of a conventional front-end). This has given good results for languages with transparent alphabetic orthographies such as Romanian, Spanish and Finnish, and can give acceptable results even for languages with less transparent orthographies, such as English [21, 22, 23, 24].

Furthermore, our tools make no use of expert-specified categories of letter and word, such as phonetic categories (vowel, nasal, approximant, etc.) and part of speech categories (noun, verb, adjective, etc.). Instead, we use features that are designed to stand in for such expert knowledge but which are derived fully automatically from the distributional analysis of plain text in the target language [21, 25].

Samples of the voices can be heard at `http://tundra.simple4all.org/demo/`. For reasons of space we refer readers interested in full presentation and evaluation of these systems to [9].

5. Conclusion

We have introduced a first version of the Simple4All Tundra corpus, and described its construction from readily available speech data. 14 audiobooks in 14 languages have been so far included in the corpus along with their orthographic transcripts. Tundra will be extended in the future with other types of imperfect, found data, such as lectures, or parliamentary speech, data which have a higher degree of spontaneity and expressivity. We will also aim at making available finer-grained alignments of the data, and also more elaborate prosodic annotations, such as style diarisation, emphasis or sentiment analysis. The TTS systems built from this corpus demonstrate a first application of the Tundra corpus, and support its usefulness.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement Nº 287678. The research presented here 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 Mihai Nae from Cartea Sonora for releasing the Romanian data, as well as to all the volunteers at Librivox and Gutenberg for dedicating their time to distribute this wide variety of data.

7. References

[1] É. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering Expressive Speech Styles in Audiobooks Using Glottal Source Parameters," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 1821–1824.

[2] N. Braunschweiler and S. Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 1821–1824.

[3] O. Boeffard, L. Charonnat, S. L. Maguer, and D. Lolive, "Towards Fully Automatic Annotation of Audio Books for TTS," in *Proceedings of LREC'12*, Istanbul, Turkey, may 2012.

[4] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *INTERSPEECH*, 2007, pp. 2901–2904.

[5] K. Prahallad and A. Black, "Segmentation of Monologues in Audio Books for Building Synthetic Voices," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 5, pp. 1444–1449, 2011.

[6] X. Anguera, N. Perez, A. Urruela, and N. Oliver, "Automatic synchronization of electronic and audio books via TTS alignment and silence filtering," in *ICME*, 2011, pp. 1–6.

[7] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. of Interspeech*, 2010, pp. 2222–2225.

[8] J. Kominek, A. W. Black, and V. Ver, "CMU Arctic Databases for Speech Synthesis," Tech. Rep., 2003.

[9] O. Watts, A. Stan, Y. Mamiya, M. Giurgiu, R. Clark, J. Yamagishi, and S. King, "Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis," 2013, in preparation.

[10] G. Bordel, M. Peñagarikano, L. J. Rodríguez-Fuentes, and A. Varona, "A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions," in *INTERSPEECH*, 2012.

[11] P. J. Moreno, C. F. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *ICSLP*, 1998.

[12] Y. Mamiya, J. Yamagishi, O. Watts, R. A. Clark, S. King, and A. Stan, "Lightly Supervised GMM VAD to use Audiobook for Speech Synthesiser," in *Proc. ICASSP (accepted)*, 2013.

[13] A. Stan, P. Bell, and S. King, "A Grapheme-based Method for Automatic Alignment of Speech and Text Data," in *Proc. IEEE Workshop on Spoken Language Technology, Miami, Florida, USA*, 2012.

[14] P. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proc. of ICASSP*, 2009, pp. 4869–4872.

[15] S. Novotney and R. M. Schwartz, "Analysis of low-resource acoustic model self-training," in *INTERSPEECH*, 2009, pp. 244–247.

[16] A. Stan, P. Bell, J. Yamagishi, and S. King, "Lightly Supervised Discriminative Training of Grapheme Models for Improved Sentence-level Alignment of Speech and Text Data ," in *Proc. of Interspeech (submitted)*, 2013.

[17] J. Ni and H. Kawai, "An Investigation of the Impact of Speech Transcript Errors on HMM Voices," in *Proc. of 7th ISCA Workshop on Speech Synthesis*, 2010, pp. 246–251.

[18] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[19] J. M. Montero and R. Barra-Chicote, "The Albayzin 2012 Speech Synthesis Evaluation (Albayzin 2012 SS) ," in *Proc. Iberspeech 2012*, 2012.

[20] A. Stan, J. Yamagishi, S. King, and M. Aylett, "The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate," *Speech Communication*, vol. 53, no. 3, pp. 442–450, 2011.

[21] O. Watts, "Unsupervised Learning for Text-to-Speech Synthesis," Ph.D. dissertation, University of Edinburgh, 2012.

[22] A. Black and A. Font Llitjos, "Unit selection without a phoneme set," in *IEEE TTS Workshop 2002*, 2002.

[23] G. Anumanchipalli, K. Prahallad, and A. Black, "Significance of early tagged contextual graphemes in grapheme based speech synthesis and recognition systems," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-April 4 2008, pp. 4645–4648.

[24] M. P. Aylett, S. King, and J. Yamagishi, "Speech Synthesis Without a Phone Inventory," in *Interspeech*, 2009, pp. 2087–2090.

[25] J. Lorenzo-Trueba, O. Watts, R. Barra-Chicote, J. Yamagishi, S. King, and J. M. Montero, "Simple4All proposals for the Albayzin Evaluations in Speech Synthesis," in *Proc. Iberspeech 2012*, 2012.