A NEW LANGUAGE AND A NEW VOICE FOR MARYTTS

Fabio Tesser, Giulio Paci, Giacomo Sommavilla, Piero Cosi
ISTC CNR - UOS Padova
Istituto di Scienze e Tecnologie della Cognizione
Consiglio Nazionale delle Ricerche - Unità Organizzativa di Supporto di Padova, Italy
[fabio.tesser, giulio.paci, giacomo.sommavilla, piero.cosi]@pd.istc.cnr.it

1. ABSTRACT

This paper describes the development of the Italian modules and the building of a new Italian female voice for the MaryTTS Text-To-Speech synthesis system. The building of new resources, such as Natural Language Processing (NLP) modules and corpus based voices for a new language in a Text To Speech system is a costly task. MaryTTS provides a number of useful tools for automatize and simplify this task.

Nowadays two state-of-the-art speech synthesis technologies are applied on modern TTS: unit selection and HMM-based synthesis. A brief introduction about the peculiar characteristic of the HMM-based speech synthesis is given in this paper; the HMM-based synthesis approach has been chosen for its higher degree of flexibility.

In the paper, the main steps necessary to built the essential NLP modules used in a TTS system using the MaryTTS tools are described. For the Italian language, more advanced NLP modules have been implemented with respect to the basic ones provided by the automatic procedures of MaryTTS.

A detailed description of the Italian MaryTTS NLP modules (such as Lexicon, LTS rules and homograph pronunciation disambiguation, numbers expansion, Part of Speech Tagger and prosodic labels prediction) has been reported here.

The paper finally illustrates the MaryTTS process necessary to select a phonetically and prosodic balanced text corpus for TTS and reports the details of the procedure used to build the first Italian MaryTTS voice with the HMM synthesis technology.

2. INTRODUCTION

The MaryTTS (Modular Architecture for Research on speech sYnthesis) TTS (Text-To-Speech) synthesis system is a flexible and modular tool for research, development and teaching in the domain of Text-To-Speech synthesis (Schröder & Trouvain, 2003).

MaryTTS[1] is an open-source project, it is written in Java and includes a number of useful tools for adding support for a new language and adding new voices. The aim of these tools is to simplify the task of building new resources for TTS, their effectiveness can be seen from the fact that when MaryTTS was born it was originally developed for the German language; nowadays it makes available voices and support for the following languages: US English, British English, German, Turkish, Russian, and Telugu.

Figure 1 shows a simple but general functional diagram of a TTS system. The support of a new language for a TTS system includes two main tasks: i) building a basic set of Natural Language Processing (NLP) components for the new language, carrying out tasks such as tokenization and phonemic transcription; ii) the creation of the voice models in the new language.

[1] https://github.com/marytts/marytts

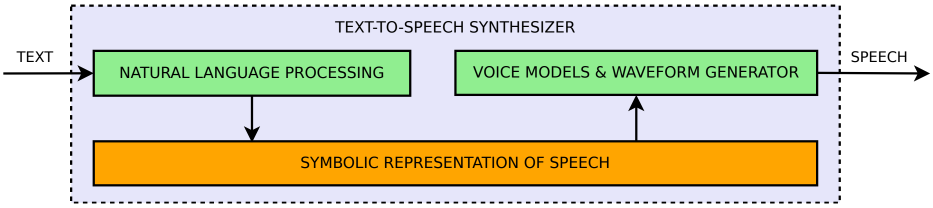
TEXT-TO-SPEECH SYNTHESIZER
TEXT
NATURAL LANGUAGE PROCESSING
VOICE MODELS & WAVEFORM GENERATOR
SPEECH
SYMBOLIC REPRESENTATION OF SPEECH

Figure 1: Functional diagram of a general TTS system.

This article describes the work carried out for adding the Italian language to the list of the languages supported by MaryTTS, and the creation of an Italian voice for the platform employing the HMM Speech Synthesis technology. The HMM Speech Synthesis approach, or more in general the Statistical Parametric Synthesis (Zen et al., 2009), has been chosen instead of the Unit Selection technology (Black & Campbell, 1995) because it allows to modify the produced acoustic patterns widely.

Statistical Parametric Synthesis takes advantage of statistical methods to generate some control parameters (usually excitation and spectral parameters) from text, and then employ them as input in a vocoder in order to generate the speech waveform.

In HMM-based speech synthesis system, the machine learning method applied is based on the HMMs theory. Figure 2 shown a functional diagram of the synthesis part of a HMM-speech synthesiser.

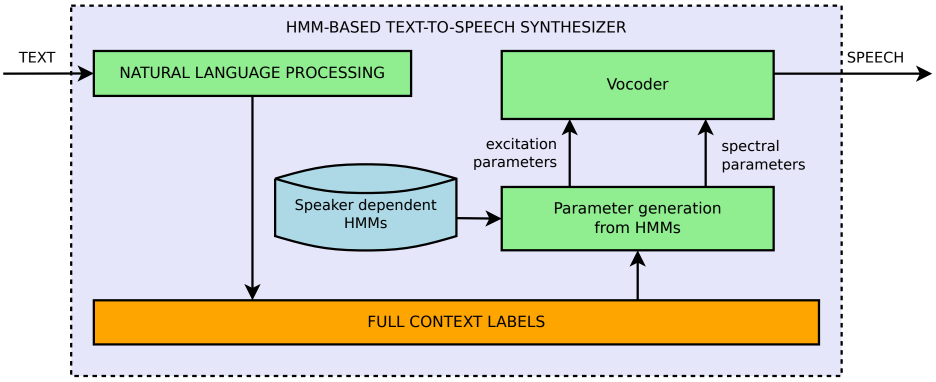

Figure 2: Functional diagram of a HMM-based TTS system.

HMM speech synthesis technology allows the functional separation of the voice models, parameters and vocoder logical blocks. For example, this technology permits to: i) stress the focus of a sentence or apply some particular prosodic patterns; ii) change the emotional content of the synthetic speech applying different prosodic settings and patterns; iii) use speaker adaptation techniques for HMM synthesis (Yamagishi et al., 2009).

HMM Italian voices have been used within the European project called ALIZ-E[2], “Adaptive Strategies for Sustainable Long-term Social Interaction”. The goal of the project is to

[2] http://www.aliz-e.org/.

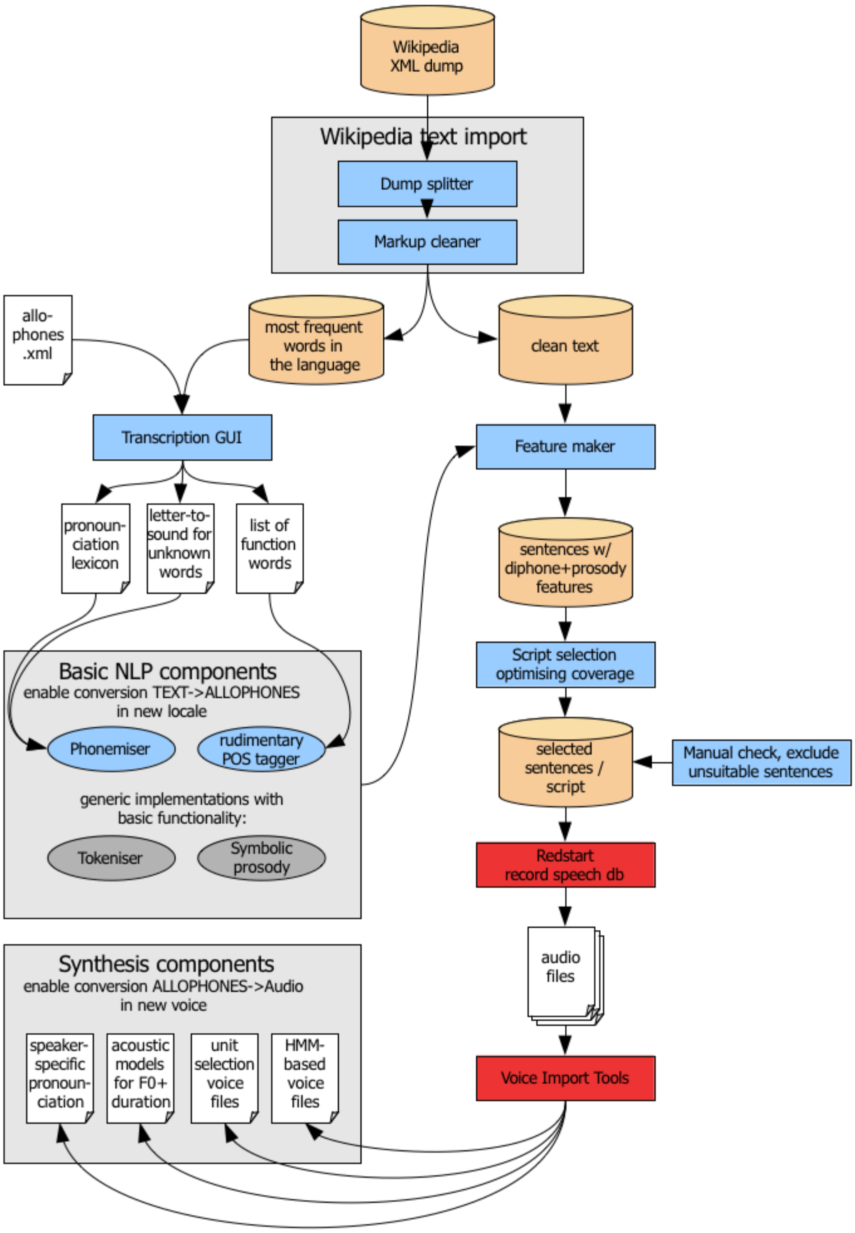
Wikipedia XML dump
Wikipedia text import
Dump splitter
Markup cleaner
allo-phones .xml
most frequent words in the language
clean text
Transcription GUI
Feature maker
pronoun-ciation lexicon
letter-to-sound for unknown words
list of function words
sentences w/ diphone+prosody features
Basic NLP components
enable conversion TEXT->ALLOPHONES in new locale
Phonemiser
rudimentary POS tagger
generic implementations with basic functionality:
Tokeniser
Symbolic prosody
Script selection optimising coverage
selected sentences / script
Manual check, exclude unsuitable sentences
Redstart record speech db
audio files
Synthesis components
enable conversion ALLOPHONES->Audio in new voice
speaker-specific pronoun-ciation
acoustic models for F0+ duration
unit selection voice files
HMM-based voice files
Voice Import Tools

Figure 3: Workflow for the *New Language Support* and *Multilingual Voice Creation* tool in MaryTTS. (Schröder & Trouvain, 2003).

develop embodied cognitive robots for believable any-depth affective interaction with young users over an extended and possibly discontinuous period.

The features of HMM-based speech synthesis systems are useful for the goals of the European project ALIZ-E, because they allow to obtain expressive voices and different voice timbres with minimal effort, using just a few speech data in the training phase.

MaryTTS supports the creation of HMM voices for new languages with the *Multilingual Voice Creation* tool. The work-flow of the *New Language Support* and *Multilingual Voice Creation* (Pammi et al., 2005) for the MaryTTS Platform is illustrated in Figure 3. The standard procedure is based on the use of the freely available Wikipedia dump of the language. The left branch of the diagram shows how to build some simple NLP modules, such as the LTS (Letter To Sound) rules for out-of-vocabulary words and a minimal POS (Part Of Speech) tagger whose unique task is to distinguish function words from content words. The right branch of the diagram is dedicated to the explanation of how to build corpus based TTS voices: from the script selection up to the real modelling of the speaker's voice using the *Voice Import Tool*.

For Italian, we decided to take as starting point some of the existing Festival TTS (Taylor et al., 1998) modules built for Italian (Cosi et al., 2001). Moreover, we developed several advanced modules to replace the basic ones provided by the standard *New Language Support* procedure.

The paper is organised as follows: Section 3 describes the process of designing and building of MaryTTS NLP modules for Italian, with particular emphasis on the modules that supersede the standard MaryTTS ones; the procedures for the creation of a new Italian female voice for MaryTTS with HMM-based synthesis technology are described in Section 4. Finally, Section 5 concludes the paper.

3. NATURAL LANGUAGE PROCESSING MODULES

As shown in Figure 1, the NLP components of a Text To Speech system are responsible for computing a symbolic representation of speech starting from the input text. There may be many levels of representation (e.g.: words, syllables, phones) to which different attributes are assigned (e.g.: POS, stress, symbolic prosody, phonetic transcription).

MaryTTS represents efficiently these levels by its own MaryXML language as shown in Listing 1.

Listing 1: MaryXML representation predicted from the input text: "Ciao mondo!".

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxml xmlns="http://mary.dfki.de/2002/MaryXML"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   version="0.5" xml:lang="it">
<p>
<s>
<phrase>
<t accent="H*" g2p_method="lexicon" ph="' tS a1 - o" pos="
   SP" pos_full="SP">
Ciao
<syllable accent="H*" ph="tS a1" stress="1">
<ph p="tS"/>
```

```
<ph p="a1"/>
</syllable>
<syllable ph="o">
<ph p="o"/>
</syllable>
</t>
<t accent="H+L*" g2p_method="lexicon" ph="' m o1 n - d o"
   pos="S" pos_full="Sms">
mondo
<syllable accent="H+L*" ph="m o1 n" stress="1">
<ph p="m"/>
<ph p="o1"/>
<ph p="n"/>
</syllable>
<syllable ph="d o">
<ph p="d"/>
<ph p="o"/>
</syllable>
</t>
<t pos="FS" pos_full="FS">
!
</t>
<boundary breakindex="5" tone="L-L%"/>
</phrase>
</s>
</p>
</maryxml>
```

3.1. Allophone set

In order to add a new language in a TTS system, the first step is that of fixing the allophones' set for the new language. The SAMPA (*SAMPA for Italian*, 1989) alphabet has been chosen because it is simple, it is a well distributed standard and because it was already used in the Italian Festival voices.

3.2. Lexicon and LTS rules

The words' pronunciation can be obtained directly from the pronunciation lexicon or, for the words not present in the lexicon, from the Letter To Sound (LTS) rules. An Italian lexicon for MaryTTS has been created, firstly converting the Italian Festival lexicon, containing around 450K words and their transcriptions, into the format specified by MaryTTS, then correcting some systematic errors and finally improving maintainability of this important resource. Afterwards the Letter To Sound rules have been inferred using an automatic MaryTTS procedure from lexicon examples.

A common problem of LTS modules is the transcription of clitics verbs[3]: these forms are in an open class, as they are derived from verb forms, and some forms lead to pronunciation

[3] In the Italian language, composed verb forms (verbs + clitics particles) are often used; these forms are composed by a base verb form plus a sequence of clitics particles (pronouns and adverbs), for example the Italian word portatemelo: `portatemelo = portate/me/lo`

ambiguities. To partially solve this problem, an automatic algorithm capable to generate clitics verbs starting from base verb forms has been implemented and used to build a lexicon with transcriptions, generating about 2.6 millions forms.

In order to treat this massive amount of data, not all the automatically generated forms are used in the Letter To Sound (LTS) training procedure: only a randomly selected part of the generated verbs is added to the base lexicon for training, with special attention to insert all the forms that may have ambiguous pronunciation. Including all the forms with ambiguous pronunciation in the training allows for automatic POS-driven disambiguation during text analysis.

3.3. Numbers expansion

Another important improvement, with respect to the basic NLP modules provided by the standard *New Language Support*, has been done in the number expansion and pronunciation. Cardinal and ordinal numbers are expanded allowing to pronounce numbers written in digit form. The expansion is a prerequisite of several expansion modules (not yet developed) such as percentages, charts, currencies, dates, that we plan to implement in the future. Cardinal numbers expansion is completed and it permits to read huge numbers, with no other limit but the size of the maximum long int. Floating point numbers are also expanded. Ordinal numbers are always expanded in the male singular form. Future work will be done to try to disambiguate ordinal numbers in order to detect gender and number.

3.4. Part of Speech Tagger

A context dependent part of speech (POS) tagger has been developed to predict whether words are nouns, verbs, or other grammatical classes depending on their surrounding context. Some manually annotated POS data for Italian has been kindly provided by (Attardi et al., 2008; Zanchetta & Baroni, 2005). This corpus contains 4000 sentences for a total of 113k words, annotated with 36 POS classes using the TANL tagset (Attardi et al., 2008). This data has been used to train an Italian OpenNLP POS tagger[4] by applying the Maximum Entropy model (Ratnaparkhi, 1997).

3.5. Homograph pronunciation disambiguation

In Italian it is common to find homograph words with different pronunciations. An example is the Italian word *ancora*, used as noun (English translation: *anchor*) or as adverb (English translation: *again*). These two Italian words have different pronunciations.

Luckily, most of the ambiguities can be solved by identifying the correct POS class for each word. To address this problem, a new data structure has been designed for the lexicon, adding the POS information; a new lexicon look-up method that make use of the POS tags has been implemented.

3.6. ToBI rules

Prosodic labels are an abstract representation of the prosody of a sentence. For example, the ToBI standard (Silverman et al., 1992) represents prosody using *break indices* that describe the degree of disjuncture between consecutive words and the tones associated with

`= portate + mi + quello`. This verb form is made up by joining the base verb form portate(=*bring, plural form*, from the verb "portare") plus me (=*to me*) to lo (=*that/it*) and it means *bring it to me*.

[4] OpenNLP library: http://opennlp.apache.org/

phrase boundaries and *pitch accents*. In TTS systems these labels are usually predicted by text-based rules (using punctuation and POS of words).

The rules provided for Italian are based on i) function words; ii) punctuation's symbols (comma, ellipsis, open brackets, colon, semicolon); iii) and sentence type (declarative, exclamatory, interrogative[5]).

4. CORPUS BASED HMM VOICE BUILDING

In order to build corpus based voice for a Text To Speech system it is necessary to design the textual corpus (script corpus). This is not a trivial task because high quality synthetic voices need to be built with a corpus that contains all possible phonetic and prosodic contexts.

Therefore, the building of a HMM-based voice for a new language require the following steps: i) script selection; ii) speaker recording; iii) HMM voice models estimation.

4.1. Script selection

The MaryTTS *New Language Support* procedure (shown in Figure 3) provides a method for optimal text selection capable of ensuring good phonetic and prosodic coverage starting from the analysis of a huge amount of text. Phonetic and prosodic information are computed over the entire Italian wikipedia dump using the Lexicon modules and the Symbolic Prosody predictor described in the previous Section. For Italian, the original sentence selection procedure has been modified in order to select only those sentences containing only words that are present in the pronunciation lexicon.

The final text selection (1400 sentences) has been obtained by 4 iterations of the following steps: a) ignore all sentences that do not improve the coverage score; b) manual inspection of the selected list and removal of those sentences that are difficult to pronounce or ambiguous; c) reiterate the coverage selection procedure.

4.2. Speaker recordings

The selected sentences have been uttered by a young Italian native female speaker (*Lucia*) and recorded in a quasi soundproof chamber. Table 1 shows the details of the *Lucia* corpus.

Feature	**Description**
Speaker	Female
Age	20
Room characteristics	Silent room
Microphone	Shure WH20QTR Dynamic Headset
DB Size (sentences)	1400
DB Size (time)	∼2 hours

Table 1: Description of the *Lucia* TTS recording corpus.

[5] Because of its different prosodic structure, interrogative sentences are classified in another category if they belong to the wh-question class. In Italian a wh-question can be detected by the first word part of speech.

4.3. HMM voice building

HMM-based speech synthesis systems make use of complex speech units called context dependent HMMs (Zen et al., 2009). These units consider the phonetic context (triphone/quinhpone models) but also prosodic and linguistic contexts such as stress, syllable accent, boundary tones, part of speech, and sentence information.

The context depended HMMs must be trained on speech[6] plus labelled data. These labels that contains all the necessary context information are called full context labels. In MaryTTS, starting from the MaryXML representation of a sentence (see Listing 1), it is possible to obtain this kind of information.

The MaryTTS *Voice Import Tools* has been used to extract full context labels and voicing strengths for mixed excitation (Yoshimura et al., 2001), while the phonetic alignment has been done using HTK 3.4.1 (Young et al., 2002).

The HTS HMM speech synthesis toolkit version 2.2 (Zen et al., 2007) has been used to build the models; mgc (mel-generalised cepstrum) spectral parameters and voicing strengths for mixed excitation are modelled using continuous probability distribution, while logF0 parts are modelled using the multi-space probability distribution (MSD) (Tokuda et al., 1999).

The *Lucia* voice has been built using the default speaker-dependent parameters of HTS: i) decision tree based state clustering; ii) separate streams to model each of the static, delta and delta-delta features; iii) single Gaussian models.

5. CONCLUSIONS

This paper has described the work done on adding Italian Language on the MaryTTS system. Advanced *Natural Language Processing* modules for TTS capable of processing textual input and extracting a symbolic representation of the speech utterance have been designed and implemented for Italian. A textual TTS corpus for Italian has been designed and a first HMM-based TTS voice has been built. As result the Italian female HMM voice *Lucia* is now available for the MaryTTS project.

At the moment no subjective evaluation has been done on the perceived quality of the voice, anyway the voice is proven to be highly intelligible and very similar to the original speaker voice. As a matter of fact, the *Lucia* HMM voice is actively employed in robot-human interactions within the EU-funded project ALIZ-E.

Further works on the Italian MaryTTS comprehend the improvement of the Italian NLP modules and the building of a male voice.

ACKNOWLEDGEMENTS

The authors wish to thank the MaryTTS team and the HTS team. Parts of the research reported in this paper have been performed in the context of the EU-FP7 project ALIZ-E (ICT-248116).

REFERENCES

Attardi, G., Montemagni, S., Simi, M., & Lenci, A. (2008). Tanl - Text Analytics and Natural Language processing: Analisi di Testi per il Semantic Web e il Question Answering (Tech. Rep.).

[6] A compact speech representation like the mel cepstrum is used.

Black, A., & Campbell, N. (1995, September), Optimising selection of units from speech databases for concatenative synthesis, in Eurospeech 1995, Madrid, Spain, 581–584.

Cosi, P., Tesser, F., Gretter, R., Avesani, C., & Macon, M. W. (2001), Festival speaks italian!, in Eurospeech 2001, Aalborg, Denmark.

Pammi, S., Charfuelan, M., & Schröder, M. (2005), Multilingual voice creation toolkit for the MARY TTS platform, in Proc. Int. Conf. Language Resources and Evaluation, Valleta, Malta.

Ratnaparkhi, A. (1997). A simple introduction to maximum entropy models for natural language processing (Tech. Rep.).

SAMPA for Italian. (1989). Retrieved on June 21, 2013, from `http://www.phon.ucl.ac.uk/home/sampa/italian.htm`

Schröder, M., & Trouvain, J. (2003), The German text-to-speech synthesis system MARY: A tool for research, development and teaching, International Journal of Speech Technology, Vol. 6, no. 4, 365–377.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., ... Hirschberg, J. (1992), ToBI: A standard for labeling English prosody, in Second International Conference on Spoken Language Processing, Banff, Canada, Vol. 2, 867–870.

Taylor, P., Black, A. W., & Caley, R. (1998), The architecture of the Festival speech synthesis system, in Proceedings of the The Third ESCA Workshop in Speech Synthesis, Jenolan Caves House, Blue Mountains, Australia, 147–151.

Tokuda, K., Masuko, T., Miyazaki, N., & Kobayashi, T. (1999), Hidden Markov models based on multi-space probability distribution for pitch pattern modeling, in Proceedings of 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, Civic Plaza, Hyatt Regency - Phoenix, Arizona, Vol. 1, 229–232.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., ... Renals, S. (2009), Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, no. 6, 1208–1230.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (2001), Mixed Excitation for HMM-based Speech Synthesis, in Eurospeech 2001, Aalborg, Denmark.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., ... Woodland, P. (2002). The HTK book (for HTK version 3.2) (Tech. Rep. No. July 2000). Cambridge University.

Zanchetta, E., & Baroni, M. (2005), Morph-it! A free corpus-based morphological resource for the Italian language, in Proceedings of Corpus Linguistics 2005, Birmingham.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., & Tokuda, K. (2007), The HMM-based speech synthesis system (HTS) version 2.0, in The 6th International Workshop on Speech Synthesis, Bonn, Germany, 294–299.

Zen, H., Tokuda, K., & Black, A. W. (2009), Statistical parametric speech synthesis, Speech Communication, Vol. 51, no. 11, 1039–1064.