Voice Controlled Child-Robot Interactions - Development of ASR and TTS systems for the NAO Robot within the ALIZ-E Project

Giacomo Sommavilla, Giulio Paci, Fabio Tesser, and Piero Cosi
ISTC-CNR, UOS Padova
Padova, Italy
{ piero.cosi, giulio.paci, giacomo.sommavilla, fabio.tesser }@pd.istc.cnr.it

***Abstract*—This paper describes the development of a voice controlled child-robot interaction system for the NAO robot platform within the ALIZ-E project. The ALIZ-E integrated system includes various components but we mainly concentrate on describing the Automatic Speech Recognition (ASR) and the Text To Speech (TTS) synthesis components and their performance.**

***Keywords*—ASR, TTS, Child-Robot Integration.**

I. Introduction

THE aim of the ALIZ-E[1] project is to develop embodied cognitive robots and to study the theory and practice of believable, any-depth and affective interactions between children and cognitive robots for an extended and possibly discontinuous period of time.

Specifically, ALIZ-E is testing robots with children (target age 8-11), who have metabolic disorders, such as diabetes or obesity. We aim for the robots to support the children's well-being and facilitate therapeutic activities in a hospital setting. The challenges of Child Robot Interaction (CRI) "in the wild" are significant with both technical and pragmatic issues to be faced, since children are not "mini-adults", and this fact is very much apparent in the context of CRI. Children bring an imaginative investment to encounters with robot agents that is hugely valuable in the exploration of how we can develop technologies and systems for social interaction. It is necessary to research their specific needs and to develop systems that address these needs, as suggested by Narayanan et al. [1] and Yildirim et al. [2]. Speech is the principal mode of communication for the child-robot interactions of the ALIZ-E project. For this reason, a significant research and development work has been dedicated to investigate and develop specific Text To Speech (TTS) and Automatic Speech Recognition (ASR) systems for the Italian language, to be used with children involved in the experimental part of the project.

A. *The* ALIZ-E *Integrated System*

The ALIZ-E integrated system is a robotic software/harware environment that implements several game-like activities that the child user can undertake. One of these activities is the Quiz Game, which is a very challenging task with regard to Child-Robot verbal interaction. The Quiz Game interaction is similar to the "Who Wants to Be a Millionaire?" game show. The child and the robot are two players who take turns asking each other questions. Whoever asks also provides multiple choice replies for the other player.

From a technical point of view, the software components of the ALIZ-E Integrated System need to (a) have access to low-level hardware devices of the Robot, (b) perform heavy computations, (c) typically run on machines connected to the Nao, (d) be coordinated concurrently and/or (e) react to (typically asynchronous) events.

Languages such as C/C++ can suit well low-level and heavy computational tasks, but it can be difficult and time consuming to manage concurrency, network communication and event handling with those programming languages. The URBI environment, developed by Aldebaran ALIZ-E project partner, provides the high level `urbiscript` scripting language which can orchestrate complex organizations of low level components called "UObjects" in highly concurrent settings.

The most interesting issues about the integration of the ASR and the TTS components into the ALIZ-E system are explained in Sections II-B and III-A, respectively.

II. Children Speech Recognition

ASR and acoustic analysis of children voices have been studied extensively by speech technology researchers. Although most of the literature comprise English speakers, in the recent years also the Italian language has been studied, and corpora have been collected and made available to the scientific community.

Lee et al. [3] have investigated spectral acoustic parameters of children speech as a function of age and gender and compared them to those of adults. This study has measured that some parameters converge to adult levels around the age of 12, while most of the acoustic speech characteristics becomes fully established around age 15.

Another important work on this matter has been carried out by Potamianos and Narayanan [4], and it is one of the first efforts to apply algorithms related to automatic recognition of children's speech. In addition to anatomical and morphological differences in the vocal tract geometry with respect to adults,

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

children have proven to introduce more disfluencies than adults, not only in spontaneous speech but also in read speech.

These are due to a not yet mature control of articulators and suprasegmental aspects of speech such as tone, stress, and prosody. Potamianos and Narayanan [4] studied the age-dependent variability of speech characteristics, due to developmental changes, that concurs to worsen ASR results when applied to children voices. An analysis on age-dependent scaling in formant frequencies, in particular the first two, F1 and F2, showed that they change almost linearly with increasing age. Also, it has proven that for children it is more difficult to classify patterns based on spectral feature as there is a high dynamic range for acoustic parameters values. The study ends with speech recognition tests with adult acoustic models on children voice inputs. Extensive experiments, conducted on Connected Digit and Command Phrase recognition tasks, showed that results get worse as age decreases. On average, ASR experiments with adult Acoustic Model tested on children lead to a Word Error Rate that is 2-5 times worse than recognition of adult AM tested on adults. In [4], applying Speaker Normalization and Frequency Warping, an improvement on ASR performance was shown.

Also Gerosa et al. [5] have studied children read speech, both for Italian and English languages, with the purpose of analyzing acoustic characteristics related to ASR. In particular, they focused on the analysis of phone duration, intra-speaker variability and acoustic space. This work also describe experiments carried out on speech recognition of children voices in matched (i.e., training and testing on voices of Italian children aged 7-13) and unmatched conditions (i.e., testing on children's speech with models trained on adult speech). It is important to note, for the benefit of the goals of the ALIZ-E project, that these tests showed that an acoustic model trained on children recordings lead to better ASR results for children inputs than a model trained on adult voices.

In [5] they have also investigated and analyzed the correspondence among vocal tract morphology, speech acoustics and formant patterns in children and in adults. Their work confirmed literature results also for the Italian language: formant frequency values of recordings of the ChildIt corpus and APASCI corpus (the former comprising children voices, the latter made up with adult speech). What is most important, with respect to the aims of the ALIZ-E project, is that Gerosa et al. [5] confirmed that 7-13 years is not an homogeneous group of speakers; in particular, for male children, the fundamental frequency and the first three formants abruptly change their values around the age of 12 years. To cope with variability of spectral parameters among different age groups, voice adaptation techniques have been tested and proven to be effectively useful to reduce errors in ASR.

During the last few years, many different Automatic Speech Recognition frameworks have been developed for research purposes and, nowadays, various open source automatic speech recognition toolkits are available to research laboratories. Systems such as HTK [6], Sonic [7], [8], Sphinx [9], [10], RWTH [11], Julius [12], Kaldi [13], the more recent ASR framework Simon [14], and the relatively new system called Bavieca [15] are a simple and probably the most famous list.

A. Julius ASR

Julius[2] has been chosen as the ASR decoder for the ALIZ-E project, mainly because it is designed for real-time decoding and modularity [12]. Its well-designed decoder API made it very easy to implement and incorporate speech recognition into the ALIZ-E integrated system. The core engine is a C library with very low system requirements. High-speed decoding is achieved with a small memory footprint. It is also possible to swap language models at run-time. Finally, Julius supports several AM normalization algorithms, including Vocal Tract Length Normalization (VTLN). For these reasons, Julius is particularly suitable for the ALIZ-E integrated system, which needs to handle several components in real-time. Its configuration is modular (i.e., each configuration file can embed another one covering only one particular aspect of the configuration).

Julius also integrates a Voice Activity Detector (VAD) based on Gaussian Mixture Models and Energy. It is very useful since it allows to avoid unnecessary coding and transmission from the ASR component. It also permits to prevent the ASR system to spot false positive results (such as words that may come from speech recognition of noise events).

B. ASR Component

In the ALIZ-E integrated system, ASR is provided as an `urbiscript` component API, whose functions can be accessed by other components (e.g., a Dialogue Manager). When an ASR output result is available, an event is launched and the result is provided as a payload, so that every components that needs this information can access it.

The ASR component basically consists of two modules. The first is a configuration structure (that holds data for Acoustic and Language Models) and a main recognition loop function (also called "Julius stream"). The second contains an internal VAD and outputs the recognition result. The principal methods of this component are: load/free/switch configuration; start/stop main recognition loop.

A diagram that describes function call and data exchange among spoken interaction components in the ALIZ-E system can be seen in Figure 1. In the Figure, Dialogue Manager (DM) and Natural Language Understanding (NLU) are two components that are connected with the ASR module. The former can specify an ASR configuration and decides when to start/stop the recognition loop. The latter takes as input the words recognized and it is responsible for interpreting them.

Julius can express its output as nbest lists or word lattices. The former is the list of n most likely sentences recognized by the ASR system. The latter, also called "Word Graph" in Julius terminology, is an acyclic ordered graph, in which nodes are words and edges are transition probabilities, weighted by the acoustic and language model probabilities.

A Word Graph is a more powerful representation than nbest lists for Spoken Language Understanding (SLU) since lattices provide a larger set of hypotheses and a more accurate representation of the hypothesis space. Both the nbest list

[2] Open-Source Large Vocabulary Continuous Speech Recognition Engine Julius: http://julius.sourceforge.jp/enindex.php

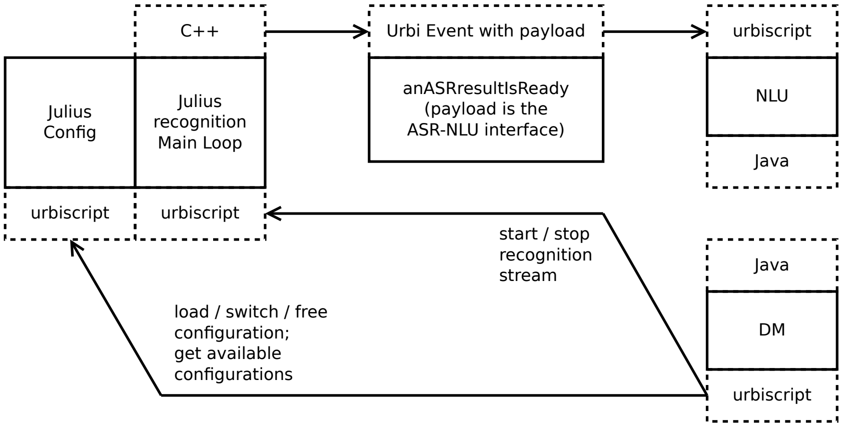
C++
Urbi Event with payload
urbiscript
Julius Config
Julius recognition Main Loop
anASRresultIsReady (payload is the ASR-NLU interface)
NLU
Java
urbiscript
urbiscript
start / stop recognition stream
Java
DM
load / switch / free configuration; get available configurations
urbiscript

Fig. 1. ASR communications through `urbiscript`.

and the lattice structures have been implemented in the Julius C++/URBI module, exposing most of the original Julius data structure to `urbiscript`.

C. Acoustic Model

The Acoustic Model (AM) for Italian children ASR has been created with the following corpora.

- The Italian FBK ChildIt Corpus [5]. The corpus is composed of Italian children voices, counting almost 10 hours of speech from 171 different children; Audio prompts consists of adults interviewing children about his/her preferred books, TV shows, hobbies, sports, etc. The audio recordings have performed using a Shure SM10A head-worn mic at 48 kHz and down-sampled at 16 kHz, 16 bit linear.
- A "read speech" corpus collected within the ALIZ-E project. It counts more than 10 hours of children speech from a total of 96 young speakers, aged 6-14.

Since the Julius distribution does not include specific tools for AM training, the HTK tools [6] have been used for this task. Also a procedure to build multi-gaussian AM has been implemented. More details on the AM training procedure can be found in the work of Paci et al. [16].

D. Language Model

Julius supports N-gram, grammar and isolated word Language Models (LM), although its distribution does not include any tool to create them. An external program must be used to create an LM.

The SRILM toolkit [17] has been used to train a 4-gram model for question recognition of the Quiz Game ALIZ-E scenario. The Quiz questions and answers database has been used as training material for a "question recognition" model. Also a simple grammar model for Quiz answers has been built, automatically including the answers in the Quiz database and then adding rules to handle common answers and filler words. Details on how the language models have been trained can be found in the work of Paci et al. [16].

E. Adaptation

As discussed in II, children voices in the age 7-13 years form an extremely heterogeneous set of spectral features. This means that the standard Hidden Markov Model (HMM) approach to ASR lead to poor results. Several works in the literature of this field suggest to apply adaptation techniques to cope with this problem.

Through the VTLN technique [6], [18], it is possible to divide the group of young speakers into more "homogeneous" groups. Also, since each child will interact several times with the robot, data from previous interactions can be used to adapt the models. For these reasons, the VTLN technique has been proven to be extremely suitable for the aims of ALIZ-E.

The procedure to apply VTLN to ASR has been the following: N recognitions, with N different configurations for VTLN parameters, with N tuned with a grid search experiment, have been run simultaneously using an AM trained with the ChildIt corpus. The configuration with the best confidence score is chosen [16].

F. ASR Results

Preliminary results of the ALIZ-E ASR system for Italian children voices have been discussed in the work of Paci et al. [16]. The test set consists of audio collected in a scenario as close as possible to the real one. For this reason it has been decided to collect and manually annotate speech data recorded during Quiz game experiments. The collected audio data consists of spontaneous speech recordings of children utterances produced during real interactions with NAO in a non-autonomous Wizard of Oz modality (i.e., recognition and understanding of user input are not automatic, but performed by a human operator). Speech has been annotated in order to provide a reference text. Also non verbal sounds (such as user's "fillers" –laughters, breath, etc.– and Nao's sounds –speech, motor noise, etc.–) have been transcribed. This represents a proper test set, because it consists of audio collected in a scenario as close as possible to the real one.

Table I shows the results of ASR applied to real-case scenario audio files, where a child poses quiz questions to

TABLE I. PRELIMINARY ASR RESULTS ON QUIZ QUESTION RECOGNITION.

ID	Snt \#	Wrd \#	WCR %	Sub %	Del %	Ins %	WER %
1	4	22	86.4	13.6	0.0	13.6	27.3
2	6	82	76.8	22.0	1.2	23.2	46.3
3	5	40	70.0	25.0	5.0	17.5	47.5
4	7	62	75.8	12.9	11.3	4.8	29.0
5	15	114	93.9	4.4	1.8	5.3	11.4
6	4	49	65.3	30.6	4.1	12.2	46.9
7	12	106	58.5	35.8	5.7	4.7	46.2
8	11	84	70.2	23.8	6.0	9.5	39.3
Total	64	559	74.6	20.9	4.5	10.2	35.6

TABLE II. PRELIMINARY ASR RESULTS ON QUIZ QUESTION RECOGNITION (VTLN).

ID	Snt \#	Wrd \#	WCR %	Sub %	Del %	Ins %	WER %
1	4	22	86.4	13.6	0.0	22.7	36.4
2	6	82	80.5	18.3	1.2	24.4	43.9
3	5	40	72.5	27.5	0.0	22.5	50.0
4	7	62	79.0	14.5	6.5	1.6	22.6
5	15	114	93.9	4.4	1.8	4.4	10.5
6	4	49	65.3	26.5	8.2	12.2	46.9
7	12	106	59.4	34.0	6.6	4.7	45.3
8	11	84	69.0	25.0	6.0	8.3	39.3
Total	64	559	75.7	20.2	4.1	10.4	34.7

the NAO robot. The 4-gram language model described in Section II-D and the acoustic model described in Section II-C have been used, without applying speaker adaptation procedures. Results of the same setup, but using VTLN techniques are shown in Table II.

The VTLN technique allows to improve all the metrics with the exception of the insertion rate (Ins). Comparing results file by file, VTLN's WCR is always equal or better than the baseline, and VTLN's WER is worse only for files 1 and 3.

Some errors are caused by Out-Of-Vocabulary words (21 errors), produced by false starts, repetitions and misreadings. 13 insertions, which are monosyllabic words, occurred for unknown reasons. The word "essere" has been spotted by the ASR system but actually corresponds to the robot talking sound. Several errors occurred when the system had to recognize sequences of monosyllabic words that are very similar. Also, in some cases errors originated by a mismatch between the actual pronunciations and the phonetic dictionary entries.

III. TEXT TO SPEECH SYNTHESIS

It is known that several messages are contained in the speech signal. Table III shows the main messages and its speech correlates. Within the ALIZ-E project, the robot has to convey to the child these messages, encoded by particular speech patterns.

TABLE III. MESSAGES CONTAINED IN THE SPEECH SIGNAL AND THEIR MAIN ACOUSTIC CORRELATES.

Message	**Acoustic correlates**
verbal content of the speech	spectral envelope
speaker's identity	spectral envelope, voice quality
emotional state of the speaker	voice quality, prosody
focus	prosody

Thus it has been required to develop a speech synthesizer capable of generating these speech patterns in order to communicate to the child those different messages. Moreover, the NAO robot can also integrate the audio/speech channel with other media: movements, gestures and blinking lights.

For these reasons it has been decided to adopt the MaryTTS system [19]. It is also released as an open source project[3].

An automatic Text To Speech system (or "engine") is usually made up by two components, a front-end and a back-end. The front-end takes care of performing Natural Language Processing (NLP) tasks. Its three main purposes are the following:

1) to normalize the input text, an operation also called "tokenization", (which means, for example, converting numbers and abbreviations into their written-out words equivalent),
2) to perform text-to-phoneme (or grapheme-to-phoneme) conversion (which means the process of assigning phonetic transcriptions to words) and
3) to divide and mark the input text into prosodic units, like phrases, clauses, and sentences.

The output of the front-end is the so called "symbolic linguistic representation", that is made up by phonetic transcriptions and prosody information. The NLP modules developed for Italian MaryTTS are described in Section 1.3.1.

The back-end – often referred to as the synthesizer or the vocoder – takes care of converting the symbolic linguistic representation into sound.

The Hidden Markov Model (HMM) Speech Synthesis approach, a declination of Statistical Parametric Synthesis [20], has been chosen for the task of modelling the voice of the robot, because it allows to widely modify the produced acoustic patterns and to provide greater flexibility than other TTS approaches, such as the Unit Selection technology [21]. In HMM systems, the back-end module needs also to compute the target prosody (pitch contour, phoneme durations), which is then imposed on the output speech.

In HMM-based speech synthesis systems, a symbolic representation of the speech segments, together with their phonetic and prosodic context, is extracted from the input text. Such representation is defined by so called "full context labels". In order to generate the speech signal, a Machine Learning algorithm uses these labels to generate the appropriate control parameters (usually excitation and spectral parameters) and then employ them as input for a vocoder. Figure 2 shows a functional diagram of the synthesis part of a HMM-speech synthesizer.

A. Integration in the Robotic Environment

A TTS component has been developed for the integrated system of the ALIZ-E project. This component is a client for the MaryTTS server.

In order to make NAO a believable robot, it is crucial that it emits the speech signal from its loudspeakers, located in its head, but this is not a trivial achievement. In fact, due to the NAO computational limitations, it has been decided to run the MaryTTS server on a remote machine, connected to the robot through a Wi-Fi network.

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

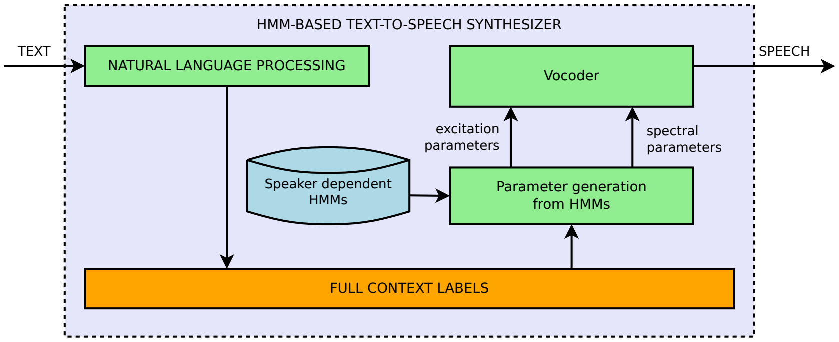

Fig. 2. Functional diagram of a HMM-based TTS system.

MaryTTS has been integrated in the `urbiscript`-based robotic system in a straightforward manner by means of an URBI UObject, called "robot.proxy.ALAudioPlayer.playWebStream", capable of directly calling the MaryTTS server and playing the speech signal using a specific http request. Moreover the low-latency of this process has been granted by the MaryTTS stream modality, which permits to stream the synthetic speech signal as soon as the first audio data are produced by the remote TTS server.

IV. Conclusions

Investigation in speech technologies is an essential part in the ALIZ-E project, since verbal interaction plays a central role in child-robot interactions. Voice controlled robots supporting hospitalized children need to incorporate adequate speech comprehension and production tools, which have to be set up for the Italian language.

The robot is needed to understand the children language, which itself is a challenging scientific task. Moreover, the scientific literature for Italian in this matter is poor if compared to that of the English language.

For these reasons, a particular care has been taken to speech recognition adaptation techniques and a big effort has been put to the collection and annotation of children speech data. Both read sentences (which allow to easily obtain transcriptions) and spontaneous utterances (taken from real case interactions) have been collected. In addition, speech data from a listen and repeat experiment have been recorded. These data will be published as speech corpora and they will be freely available to the scientific community.

On the other hand, the robotic verbal output, applied to interactions with hospitalized children, should convey expressivity and emotions, in order to involve and engage the young users as much as possible. Finally, stress on particular words or phrases is crucial for the children to comprehend the most important educational topics of the interactions. Specific Text to Speech modules and tools for the Italian language have been studied and developed.

Acknowledgment

This work was partially supported by the EU FP7 ALIZ-E project (grant number 248116).

References

[1] S. S. Narayanan and A. Potamianos, "Creating conversational interfaces for children." *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, 2002.

[2] S. Yildirim, S. Narayanan, and A. Potamianos, "Detecting emotional state of a child in a conversational computer game," *Comput. Speech Lang.*, vol. 25, no. 1, pp. 29–44, Jan. 2011.

[3] S. Lee, A. Potamianos, and S. S. Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999, selected Research Article.

[4] A. Potamianos and S. S. Narayanan, "Robust recognition of children's speech." *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 2003.

[5] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, pp. 847–860, Feb 2007.

[6] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[7] B. L. Pellom and K. Hacioglu, "SONIC: The university of colorado continuous speech recognizer TR-CSLR-2001-01," University of Colorado, Boulder, Colorado, Tech. Rep., March 2001.

[8] ——, "Recent improvements in the CU Sonic ASR system for noisy speech: the SPINE task," in *ICASSP (1)*, 2003, pp. 4–7.

[9] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 1, pp. 35 – 45, January 1990.

[10] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Sun Microsystems, Inc., Mountain View, CA, USA, Tech. Rep., 2004.

[11] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, "The rwth aachen university open source speech recognition system," in *Interspeech*, Brighton, UK, Sep. 2009, pp. 2111–2114.

[12] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, 2001, pp. 1691–1694.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[14] "Simon listens webpage," http://www.simon-listens.com, accessed: 2014-06-20.

[15] D. Bolanos, “The bavieca open-source speech recognition toolkit.” in *SLT*. IEEE, 2012, pp. 354–359.

[16] G. Paci, G. Sommavilla, F. Tesser, and P. Cosi, “Julius ASR for Italian children speech,” in *9th national congress, AISV (Associazione Italiana di Scienze della Voce)*, Venice, Italy, 2013.

[17] A. Stolcke, “SRILM - an extensible language modeling toolkit,” in *Proceedings of eventh International Conference on Spoken Language Processing (ICSLP)*. ISCA, 2002, pp. 901–904.

[18] P. Zhan and A. Waibel, “Vocal tract length normalization for large vocabulary continuous speech recognition,” CMU Computer Science Technical Reports, Tech. Rep., 1997.

[19] M. Schröder and J. Trouvain, “The German text-to-speech synthesis system MARY: A tool for research, development and teaching,” *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[20] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[21] A. Black and N. Campbell, “Optimising selection of units from speech databases for concatenative synthesis,” in *Eurospeech 1995*, no. September, Madrid, Spain, September 1995, pp. 581–584.