

OLA – Organizational Life Assistant

FOR FUTURE ACTIVE AGEING

D2.2 Spoken dialog interaction

Project Identification	
Project Number	AAL 2014-076
Duration	38 months (1 st March 2015 – 30th April 2018)
Coordinator	Carla Santos
Coordinator Organization	Inovamais, S.A. (INOVA+)
Website	http://project-ola.eu/

Document Identification	
Deliverable ID	D2.2 Spoken dialog interaction
Version/Date	V1.1 / 20.09.2016
Leader of the Deliverable	ISCTE-IUL
Work Status	Finished
Review Status	Accepted


Deliverable Information	
Deliverable Description	Development of speech input and output interfaces for the OLA (virtual) spoken dialog system. With a speaker independent automatic Speech Recognition (SR) engines and enhanced Text-To-Speech (TTS) solutions for Swedish, Polish and European Portuguese.
Dissemination Level	Public
Deliverable Type	Report
Original Due Date	M16

Authorship & Review Information	
Editor	Lázaro Ourique (ISCTE-IUL) Miguel Sales Dias (ISCTE-IUL)
Partners Contributing	Morgan Fredriksson (LM)
Reviewed by	Bruno Coelho, Marco Duarte (INOVA+)



Table of Contents

1	Executive Summary.....	5
2	Document Context.....	6
2.1	Role of the Deliverable	6
2.2	Relationship to other Project Deliverables	6
2.3	Target Audience of the Deliverable	6
3	Project Description.....	7
3.1	General Description	7
3.2	System Description.....	8
3.3	Status and Future Developments	9
4	Introduction.....	10
4.1	Basic support to natural language understanding.....	10
5	Speech Data.....	12
5.1	Desktop data collection of elderly speech	12
5.2	Speech Collection Results	13
5.3	Transcription and Annotation	13
6	Development of Specialized Acoustic Models.....	15
6.1	New acoustic models for Polish and Portuguese.....	15
6.1.1	Deep belief Neural Network -based Acoustic Model.....	16
6.2	Evaluation Results of the Portuguese and Polish ASR	16
7	Personalized Text-to-Speech Voices	18
7.1	Microsoft Hidden Markov Models-based Text-to-Speech system (HTS) System for Portuguese and Polish.....	18
7.1.1	General view of Synthesis.....	19
7.1.2	High quality Voice font creation	19
7.1.3	Training models (SPS and VA)	20
7.2	Swedish HMM-based TTS.....	23
7.2.1	Speech databases.....	24



7.2.2	Context dependent labelling and decision trees	25
7.2.3	Novel excitation model for Swedish HMM-TTS	26
7.2.4	Codebook-based excitation model	27
7.3	Final version of the TTS voices in Swedish, Polish and Portuguese.....	30
8	SDK and language packs	31
9	Conclusion	32
	List of Figures.....	33
	References	34



1 Executive Summary

In the context of the OLA project a special focus was placed in the development of Speech output and input interfaces, as the partners previous experience revealed this was of great importance for the easy of navigation in apps by elder users.

Two systems were envisioned an Automatic Speech Recognition (ASR) and an Text-To-Speech Synthesis (TTS) in the target market languages, Portuguese (pt-PT), Polish (pl-PL) and Swedish (sw-SW). For the pt-PT and pl-PL the group has focus its efforts in the development of especially adapted systems for the elder user, since ASR and TTS language packages have already been developed by technological companies such as Microsoft and are freely and publicly available to developers worldwide. As for sw-SW the group will focus on the development of a new language package since this language is not currently available.

2 Document Context

2.1 Role of the Deliverable

This deliverable presents the current status of development of the new Speech Recognition (SR) and Text-to-Speech (TTS) tools being developed for the OLA project, as well as, introducing the various concepts, methodologies and techniques deployed in the development of this technologies.

2.2 Relationship to other Project Deliverables

Deliv.	Relation
D1.1	Title: User requirements specification and use case definition D1.1 defined the requirements for the Speech Recognition (SR) and Text-To-Speech (TTS) tools.
D1.2	Title: Concept development D1.2 further specified the Speech Recognition (SR) and Text-To-Speech (TTS) tools.
D1.3	Title: Technical Specification D1.3 specifies the relationship between Speech Recognition (SR) and Text-To-Speech (TTS) modules and the overall system architecture.
D2.3	Title: Multimodal interaction The Speech Recognition (SR) and Text-To-Speech (TTS) tools of the D2.2 are essential Input and Output interfaces of the OLA app.
D3.1	Title: Design specification and integrated architecture D3.1 further specifies the relationship between Speech Recognition (SR) and Text-To-Speech (TTS) modules and the overall system architecture with a special focus on software development.

2.3 Target Audience of the Deliverable

This document is a public deliverable. Still, it is mainly intended for the project partners and the European Commission services thus the document will be made public, but not specifically disseminated on a wider scale.

3 Project Description

3.1 General Description

This project aims to offer an answer to the societal challenges by providing an innovative Organizational Life Assistant (OLA), a virtual presence that supports instrumental activities relating to daily living needs of older adults allowing them to be more independent, self-assured and to have a healthier, safer and organized life, while easing caregivers work.

OLA will mediate and facilitate interaction (communication and collaboration) between senior citizens and their informal caregivers or other services or professionals, through technological devices such as standard computers, mobile devices (tablets) and home automation modules. These ICT (Information and Communications Technology) devices will be based on an innovative multimodal model, embracing various physical/healthy and cognitive characteristics of the older adults and will be specifically oriented to increase the level of independence of the elderly, by supporting the possibility of carers' assistance remotely and by improving the accessibility to existing services on the Web, such as on-line shopping services.

Moreover, the OLA will also provide personalized well-being and safety advices to older users in order to avoid unwanted age related health and safety situations in their own home. Such a well-being and safety advisor makes uses of a combination of user information that is collected (personal physical/health and cognitive characteristics) and extracted through emotion recognition and various sensors.


OLA also addresses a major issue that elderly face related to memory degradation and gradual decreasing of their cognitive capabilities, enabling them to remember primary health care and fiscal obligations (e.g. personal hygiene, medical and tax compliance) or helping them to find everyday items such as eyeglasses, wallet or keys. It is based on speech dialogue interfaces and space and object reconstruction and classification to capture and store daily routines and their related contexts.

The primary end-users are the big group of 65+ adults living alone with or without light physical or cognitive age related limitations, who need support from care systems. Secondary end-users are both formal and informal caregivers from public or private sectors, supporting them to cope with the increased demand for care..

3.2 System Description

OLA addresses specifically the following main issues:

- **Well-being advisor:** based on the combination of the collected user information (personal, healthy characteristics) and user interaction information extracted through emotion recognition, sensors settings and contextual recorder capturing the routines as done by the older adult) the system will propose to the older adults' personal advice adapted to their situation contributing to their preservation and well-being status in home environment. In case of risk (e.g. irregular heart rate, extreme fatigue) the system may ensure an alert to a local medical emergency service.
- **Collaborative care organizer:** based on the ISCTE-IUL and LM knowledge of developing human-computer interaction platforms (HCI), OLA will provide online care collaboration between family and professional caregivers, by enabling a local care network to communicate, access sensor data, and coordinate care tasks. With the OLA assistant, seniors will be able to actively participate in the care organization through voice, even when they are unwilling or unable to use traditional web applications.
- **Safety advisor:** based on the combination of collected user environment information through real-time analysis and augmented reality settings, the system will propose suggestions of environment changes that interfere with accessible paths and provide alerts for intruders or other situations that can create hazard situations. In case of risk (e.g. checking intruders or fire), the system may contact local emergency services.
- **Every day instrumental daily living activities memory support:** the system will anticipate medical and fiscal compliances, remember primary health care and food requirements and could help elderly to find displaced everyday items.
- **Environment analysis:** algorithms for real-time object recognition and scene understanding will be developed based on a number of inputs (i.e. 3D object and space reconstruction by using time-of-flight and augmented reality technology) in order to analyse and decide which action to be taken in order support the elderly by suggesting environment changes and providing hints/advice for safety and accessible environments.
- **Multimodal interaction for elderly:** An adaptive organizational life assistant, a virtual presence will be developed in order to facilitating communication and collaboration between older-adults and informal caregivers or other services or professionals. This will be a user-friendly system that uses multimodal approaches based on non-invasive



and minimally obtrusive technologies (i.e. speech, silent speech, touch, gestures, RGB-D sensors).

The overall OLA system will be an easy to download and install software making use of multimodal integrated settings. OLA is in essence a service that enables the elderly user to reduce the demand of care through prevention and self-management, while at the same time also facilitating the supply of formal and informal care assistance.

A series of well-selected use cases where older adults have been supported by caregivers and care professional services will be developed, as well as pilots representing different use cases. Care units will use the system over a one year period. A new evaluation approach will be used during the pilots, investigating up to which point the OLA services alleviate caregivers support and maintain, or even improve the self-management, health and safe lifestyle of the older adult at home.

3.3 Status and Future Developments

Currently this deliverable sits between both versions that are planned for it (version A and B), for which the first version was already completed with some delay; however, the current stage of developments lets foresee that the final version will be achieved without any type of problems given the completion rate of the technology.

4 Introduction

The development of the ASR and TTS tools for the OLA project is mainly based on publicly available technologies from Microsoft, due to the initial design of the task and system for this module being created while Microsoft was still a consortium partner. The task and system were re-evaluated by the partners, upon Microsoft exit, resulting in some minor adjustments.


4.1 Basic support to natural language understanding

Speech technology (Automatic Speech Recognition - ASR and Text-To-Speech Synthesis - TTS), is an alternative modality of human-computer interaction and can sometimes be easier and more intuitive for the user, since speech is considered to be the most natural way of interaction for humans. For this project, ASR and TTS speech technologies were elected as the preferred way for human-machine interaction and were assessed for groups of elderly people who might have some level of body movement constraints. For instance, an Automatic Speech Recognition (ASR) engine optimised for the elderly allows elderly speech to be used as input to control applications.

There are several ASR engines that can be used for controlling computer applications. However, the acoustic models (statistical models of speech sounds) used in those engines are built using the voices of young and middle-aged adults. As people's voices change over the course of the aging process, such models typically do not work well with elderly speech and result in unacceptably high word error rates in speech-enabled applications. Since OLA is a project that specifically targets elderly users, the solution that offers the best possible speech experience was to develop new ASR engines and TTS voices specially optimized for the idiosyncrasies of elderly speech.

This project took into consideration and developed improvements to the currently available speech technologies for the Portuguese, Swedish and Polish, the countries where end-users pilot applications and usability evaluation studies will be executed.

As for TTS, new high-quality male voices were developed for the target languages, since Female voices are already available in the currently existing technology and can be integrated. Additionally, we introduced technology to enable the rapid and low-cost creation of personalized TTS voices (e.g. voices of the family members of the elderly persons), for the 3 mentioned languages. The need for these new high-quality male and personalized voices comes from the fact that we must provide a solution to the end-users' preferences for a synthetic voice.



Regarding ASR, we improved the performance for all the project languages by adapting and expanding currently available ASR systems in order to cope with speech produced by elderly people, effectively enabling this technology to be integrated into assistive prototypes adapted to the end-users [25].

One final achievement was to make available the developed TTS and ASR systems in language packs for (industry) standard SDKs, for server side (UCMA) and client side (SAPI) speech development.

5 Speech Data

5.1 Desktop data collection of elderly speech

The first step for creating the new elderly ASR engines was to collect data from elderly adults. To make the process easier and more agile, the Yourspeech platform was used. Yourspeech is a platform for speech data collection created by MLDC - Microsoft Language Development Center – that allows the recruited speakers to donate speech in their native language. Figure 1 shows the general architecture of the system.

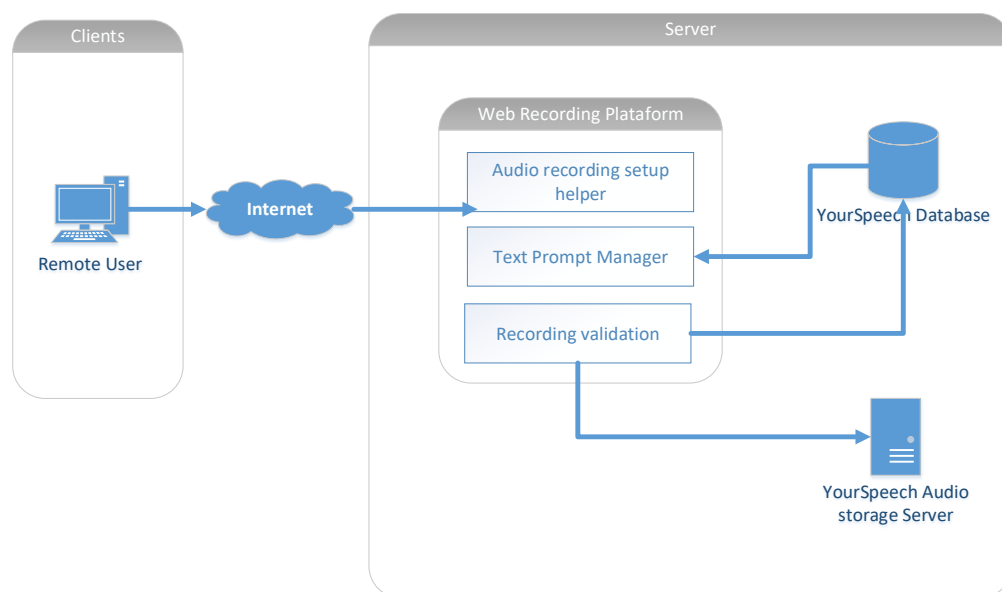


FIGURE 1 - YOURSPEECH ARCHITECTURE

Recruited speakers start by accessing the Yourspeech recording platform web page. An instance of the recording platform was created for each country and localized according to the native language. After logging in, the speakers go through the usual microphone setup phase, where microphone volume is calibrated to the appropriate level.

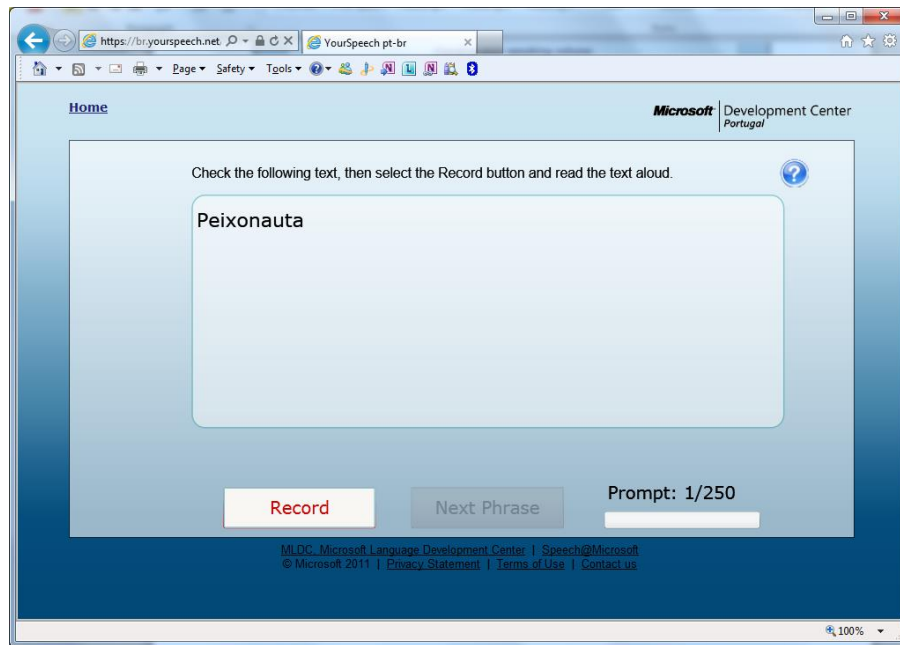


FIGURE 2 - RECORDING OF PROMPTS USING YOURSPEECH

When this phase is completed, speakers are presented with the prompts to be read (as shown in Figure 2). About half of the prompts are extracted from a pool of phonetically rich sentences that were previously produced, verified and loaded into the database. The other half of prompts are command and control type of prompts typically used in speech-enabled applications (e.g. dates, times, commands to operate applications).

5.2 Speech Collection Results


Table 1 summarizes the collected speech during the campaign in Poland, Portugal and Sweden and also shows the previously existing speech for Portugal [26]. Two values are presented: the total hours of audio recorded and the total hours of pure speech recorded (the audio excluding pauses, silences and other noises).

	Poland	Portugal	Sweden
Total audio recorded (hh:mm:ss)	203:17:23	185:10:25	~10:30:00*
Pure speech recorded (hh:mm:ss)	87:10:39	87:25:01	~4:00:00*

TABLE 1 - SPEECH COLLECTED IN THE OLA COUNTRIES. *CURRENTLY BEING PROCESSED

5.3 Transcription and Annotation

The collected speech of all the target languages was transcribed and annotated using an internal procedure specially designed for speech corpora transcription and



annotation [26]. Once the data collection finished, we ensured that the audio recorded matches the prompts presented to the speakers and made any necessary corrections to the transcriptions. In addition, using an annotation scheme, we marked the presence of noises and other audio events in the recordings. The manual verification, correction and annotation of the transcriptions is necessary because it improves the ASR engine training accuracy.

6 Development of Specialized Acoustic Models

The Automatic Speech Recognition (ASR) functionality of the OLA pilot application (OLA) is implemented using the Microsoft Speech Platform Runtime (Version 11), which contains a Hidden Markov Model (HMM) -based speech recogniser and a language pack, incorporating the language-specific components necessary for ASR: grammars, a pronunciation lexicon, and the acoustic models (AMs). The goals of the ASR-related work in the project are to create a new Swedish (sw-SW) language pack, as well as, Polish (pl-PL) and Portuguese (pt-PT) language packs that are specifically adapted to the elderly users of the OLA application, and to obtain the best possible recognition performance using existing techniques and tools compatible with the requirements of the Microsoft Speech Platform Runtime. The following subsections provide information about our acoustic models (AMs), which have been optimised for elderly speech by updating AMs trained with young to middle-aged adults' speech with the elderly speech collected during the OLA project. We omit detailed information about the original AMs trained with young to middle-aged adults' speech and the training techniques used before the project, as it is commercially-sensitive information.

6.1 New acoustic models for Polish and Portuguese

The pl-PL and pt-PT AMs that we adapted to elderly speech originate from the pl-PL and pt-PT language packs that can be used with the Microsoft Speech Platform Runtime, respectively. They comprise a mix of gender-dependent (GD) whole-word models and cross-word triphones trained using several hundred hours of read and spontaneous speech collected from young to middle-aged adult speakers. They also include a silence model, a hesitation model for modelling filled pauses, and a noise model for modelling human and non-human noises.

The pl-PL and pt-PT elderly speech corpora were divided into three datasets (training set: 85% of the speakers; development test set used for optimisation purposes: 5% of the speakers; evaluation test set used for measuring the final performance of the AMs: 10% of the speakers). The AMs included in the pl-PL and pt-PT language packs were optimised for elderly speech by retraining them with the data in the training sets. The hesitation and noise models were retrained using the stretches of audio signal that correspond to the hesitation and noise tags inserted into the transcriptions during the transcription and

annotation phase. In the case of pl-PL and pt-PT there was a total of 165.8 and 147.5 hours of audio in the training set, respectively.

6.1.1 Deep belief Neural Network -based Acoustic Model

Additionally, for the pt-PT speech data we tested a new acoustic model paradigm with preliminary support from the Microsoft Speech Platform. This paradigm based on Deep Belief Neural Networks (DNN) has been recently applied to speech recognition and is currently the state-of-the-art with ample evidence in the literature showing significant performance gains when compared with more classic approaches such as the GMM – based AMs [1, 2, 3, 4]. For this experiment we started with an existing Portuguese DNN acoustic model and adapted it using the pt-PT elderly speech training data.

6.2 Evaluation Results of the Portuguese and Polish ASR

To illustrate the improvements in ASR performance that can be achieved by using acoustic models optimized for the elderly, we trained comparable bigram language models (LM) for Portuguese (pt-PT) and Polish (pl-PL). The pl-PL corpora contain sentences extracted from newspapers. The pt-PT corpus, however, also contains command & control type of material; only about half of the utterances in the corpus correspond to sentences extracted from newspapers.


To keep the ASR results comparable across the languages, we used the sentences in the pl-PL training set – excluding the sentences that also appear in our test sets – to train the bigrams but, in the case of pt-PT, additionally excluded the command & control type of material from the training set. To compensate for the loss of training sentences in the case of pt-PT, we appended the pt-PT training set with newspaper sentences from a small in-house corpus.

The pl-PL and pt-PT test sets contained 83135 and 52970 word tokens, respectively. Table 2 summarizes the key details of the bigram language models, as well as the results and improvements gained with the specialized acoustic models as compared with acoustic models trained with young to middle-aged adults' speech (Baseline).

Language	ASR vocabulary	LM perplexity (test set)	OOV words (test set)	WER (%) Baseline AMs	WER (%) Specialised AMs	WERR (%)
Polish	19781 words	54.3	283	16.0	13.6	15.0
Portuguese	8934 words	60.0	219	18.3	16.4	10.4

TABLE 2 - ASR RESULTS WITH SPECIALIZED ACOUSTIC MODELS (AMs).

The analysis (Table 2 - ASR Results with Specialized Acoustic Models (AMs).) was based on: the Language Model (LM) Perplexity is an information theory derived measure of how



well a probability model (the LM) predicts a sample (the test set utterances); the Out Of Vocabulary (OOV) rate of the test set; Word Error Rate (WER) figures obtained on the test set by the baseline and the specialized AMs; and the Word Error Rate Relative Reduction (WERR).

The specialized AMs provided a significant improvement in ASR performance with respect to the baseline models. Furthermore, the Portuguese results show a lower performance improvement compared with Polish. This can be explained by the higher Out of Vocabulary (OOV) rate of the test set, partially masking the AM adaptation gains. As expected, the DNN modelling is capable of achieving higher performance and higher relative WER reduction when compared with the GMM-based AMs.

7 Personalized Text-to-Speech Voices

A Text to Speech – TTS – service enables text to be read by a synthesized artificial voice. One of the objectives of the OLA project was to enable communication via voice as a service more appealing to the user. We wanted to build applications that are not only easy to use, but also friendly and with a bigger human component. As voice simulates presence, it enables the app to communicate with the user in the same way the user communicates with his/her family and friends: by using speech. This gives a new meaning to “assistant applications for the elderly”.

In the beginning of the project, several TTS voices were available to be used. The resources provided by Microsoft, for example, supports female voices for Polish and Portuguese that are available for public download in this link: <http://www.microsoft.com/en-us/download/details.aspx?id=27224>. LM also created several voices for Swedish. Nevertheless, all these voice were generated based on young adults. From the know-how acquired by the various partners in previous projects, in particular ISCTE-IUL in the AAL4ALL project, elder users feel more related to and preferred elder voices. Based on this, the goal was to create at least one elderly TTS voice for each country of the project.

Since the Microsoft TTS service has no support for the Swedish language, the project couldn't use it to generate personalized TTS Swedish voices. This section present two architectures used for TTS voice generation on the OLA project. The Microsoft Hidden Markov Models-based Text-to-Speech system (HTS) System [5], and the Swedish HMM-based TTS. The Microsoft HTS system was used to create the voices in Polish and Portuguese, since these voices and natively supported by Microsoft and their products already available in several households and development ecosystems. To generate Swedish TTS voices, the Swedish HMM-based TTS was used.

7.1 Microsoft Hidden Markov Models-based Text-to-Speech system (HTS) System for Portuguese and Polish

The Microsoft HTS System was used in previous projects to generate Portuguese voices based on young adults. In OLA, it was used to generate the elderly voices for Poland and Portuguese. This section presents and describes the major parts of this system. More

specifically, the training models used are dubbed Statistical Parameter Synthesis (SPS), for Portuguese and Polish, and Voice Adaptation (VA), both based on HTS technology.

7.1.1 General view of Synthesis

The HTS synthesizer has, as most of speech synthesizers, two main parts: front-end and back-end.

The front-end is dictionary-based, being composed by a lexicon with around 135000 words, phonetically annotated by a professional linguist with phonetic transcriptions, stress marks and syllable boundaries, and with Part-of-Speech (POS) information. The stress and syllable marking was automatically assigned using linguistic rule-based algorithms, specially developed for the European Portuguese language.

The front-end is also composed by the text analysis module, which involves the sentence separator and word breaker components, including several other files, such as phone set and features and the POS tags set. It also includes a rule-based TN (Text Normalization) module [5], as well as a homograph ambiguity (also polyphony) resolution algorithm [5, 6], a stochastic-based LTS (Letter-to-Sound) converter, used to predict phonetic transcriptions for out-of-vocabulary words and the prosody models, which are data driven using a prosody tagged corpus of 2000 sentences and a POS tagger who provides morpho-syntactic contextual information.

The front-end outputs phonetic transcriptions that are subsequently input of the TTS runtime engine or back-end, which then outputs synthetic voice.

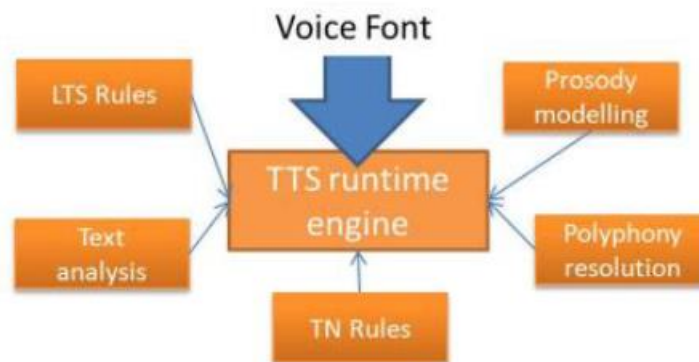


FIGURE 3 - SYNTHETIC VOICE GENERATION

7.1.2 High quality Voice font creation

The voice font building is also a very complex and demanding process that requires the following major steps):

- script selection (using different text genders, phonetically balanced, with a total of 11 500 prompts and nearly 13 hours of speech);
- recording process according to TC-Star's specifications (2 channels, audio and EGG (Electroglottograph) signal, at 96 kHz, 24 bits of sampling rate) [7];
- edition of the prompts, recording quality control, re-recording and edition of the prompts which failed in the quality control, wave process, automatic alignment and quality validation, font compiling and conversion of the original recorded waves to 8khz, 8-bits sample rate.

Figure 4 depicts the voice font building process in more detail.

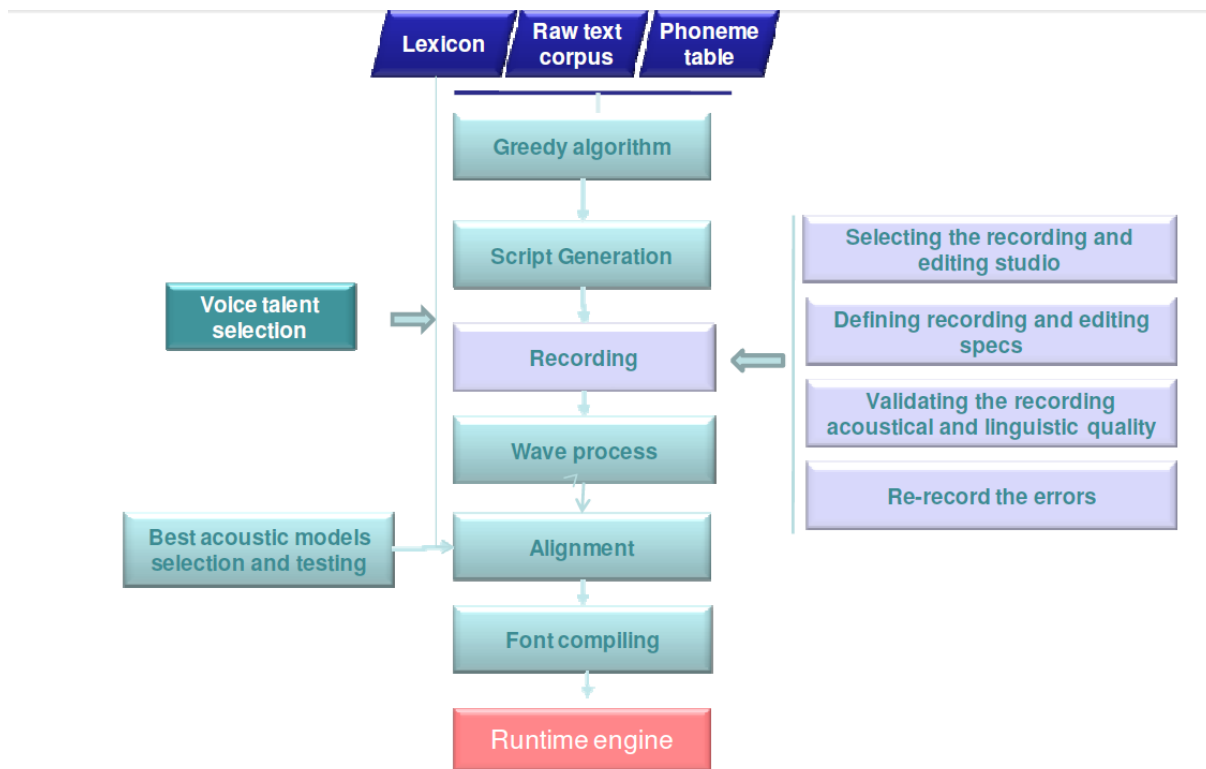


FIGURE 4 - VOICE FONT BUILDING.

7.1.3 Training models (SPS and VA)

Statistical Parametric Synthesis (SPS) requires the extraction of parameters characterizing the vocal tract emitting the relevant phone to be modelled (24 in this case). The wave segmentation (at sentence, word, syllable and phone level) is guided by the fully normalized prompts used for eliciting the recording. This procedure is automatically

carried out using rule-based sentence breakers, contextual text normalization (TN) rules, and POS taggers.

Once single words are normalized and categorized, the correct pronunciation is retrieved from the lexicon and assigned to the current word. In the end, each prompt is enriched with several types of information, as shown below (“w” = token, “v” = written form, “p” = pronunciation etc.):

```
<text>Mas na ausência destas condições de juízo desapaixonado .</text>

<words>

  <w v="Mas" p="m aex sh" type="normal" pos="conjc" regularText="mas" length="3" />

  <w v="na" p="n aex" type="normal" pos="contr" regularText="na" offset="4" length="2" />
  <w v="ausência" p="aw - z en 1 - s j aex" type="normal" pos="noun" regularText="ausência" offset="7"
    length="8" />

  <w v="destas" p="d eh 1 sh - t aex sh" type="normal" pos="contr" regularText="destas" offset="16"
length="6" />

  <w v="condições" p="k on - d i - s onjn 1 sh" type="normal" pos="noun" regularText="condições"
offset="23" length="9" />

  <w v="de" p="d ax" type="normal" pos="prp" regularText="de" offset="33" length="2" />

  <w v="juízo" p="zh w i 1 - z u" type="normal" pos="noun" regularText="juízo" offset="36" length="5" />

  <w v="desapaixonado" p="d ax - z aex - p aj - sh u - n a 1 - d u" type="normal" pos="adj" br="3"
tobifbt="L-L%" regularText="desapaixonado" offset="42" length="13" />

  <w v="." type="punc" pos="symbol" regularText="." offset="90" length="1" />

</words>

</sent>
```

The basic idea is that the waveform is stable during short time phrases and can be approximated by Gaussian models that represent the parameter distribution. Given a sequence of observation (O_t), we expect an observation O_i at time i to belong to one state Q (e.g. 1, 2 or 3). In $i+1$, O_{i+1} might still be Q or a different state and this must be modelled depending on the transition probability built on previous observations, based on the aligned wave – the prompt pairs we used for training. We used a decision tree based on relevant distinctive features associated to a given state (expressed by

Gaussian models), to better estimate the state sequence (as well as its duration and excitation).

Notice that the SPS procedure not only allows using distinctive phonetic features (line spectrum pair, LSP model, [8]) as parameters, but also prosodic cues, like pitch (F0). This allows us to both keep the advantages of having an HTS Voice Font (high flexibility, small font size) and to limit its disadvantages (muffled voice quality, flat prosody). The training process uses a gradient descent algorithm (Minimum Generation Error, MGE [9]).

In the end, the (trained) decision tree is used in generation to select and concatenate the state models by maximizing the likelihood of the parameter sequence. The result is a fully intelligible TTS voice font.

Both the training and the synthesis pipelines are depicted schematically in Figure 5.

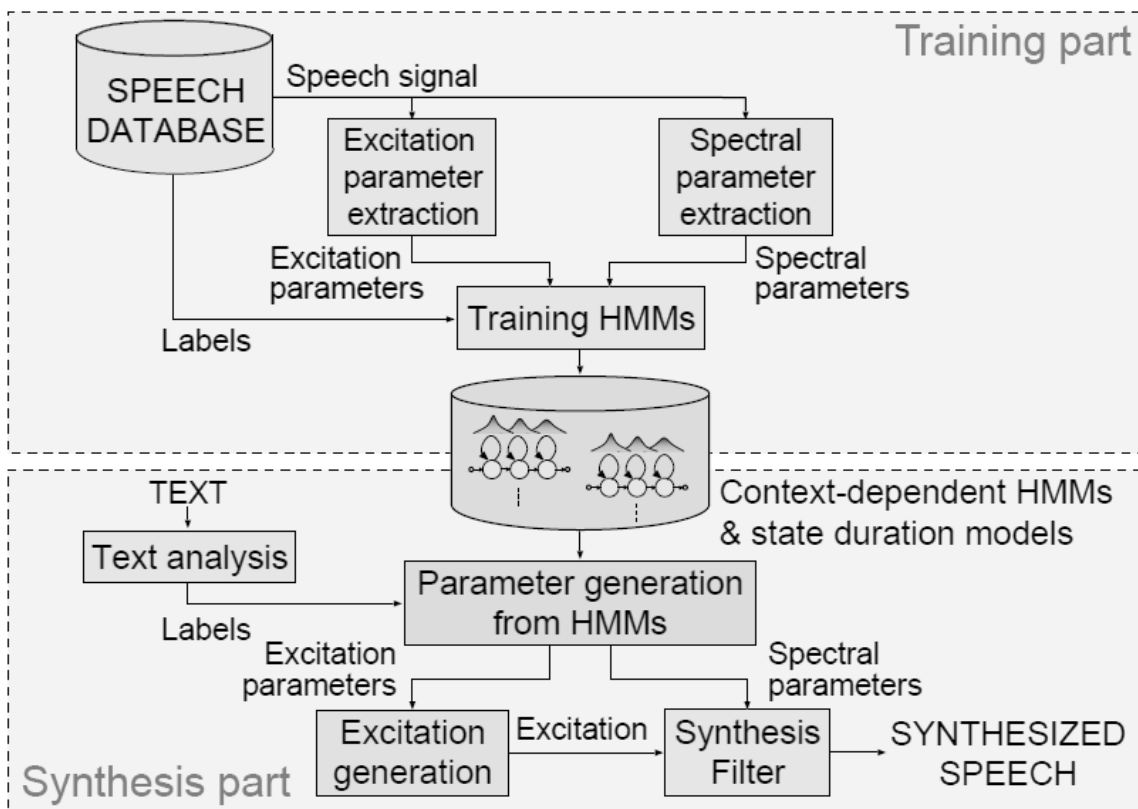


FIGURE 5 - ARCHITECTURE OF SPS-BASED VOICE FONT TRAINING AND SPEECH SYNTHESIS

While the SPS training procedure might require up to 8.000 recordings to come up with a very good voice font, Voice Adaptation (VA) could require less than 500 recordings to

obtain a voice font of similar quality. The idea behind VA is to modify a source TTS voice to sound like a target speaker by learning from a small amount of speech data recorded from the target speaker, then expand one style to other styles with low cost (Yes-No Question, Navigation, Commands etc.). Once a Source Speaker's HMM models are present, VA is possible: the decision tree will be kept the same, while only the model parameters will be transformed as schematized in Figure 6.

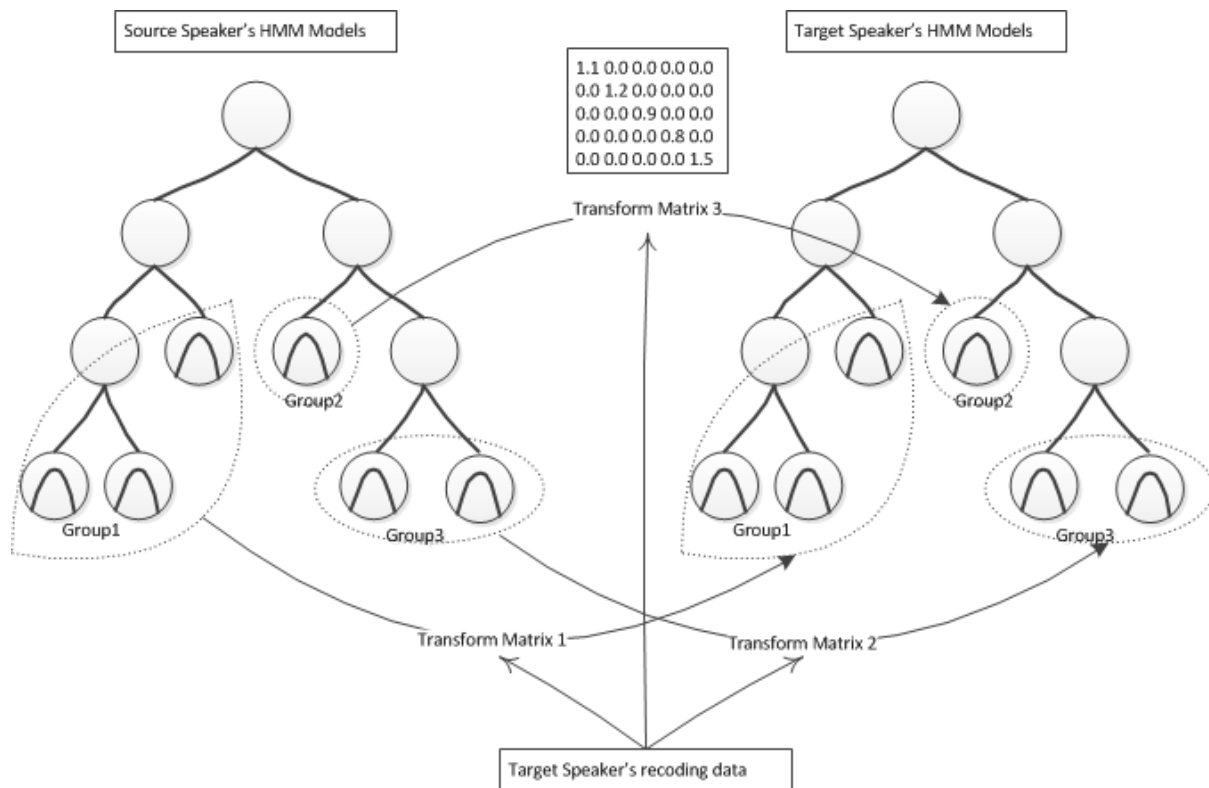


FIGURE 6 - VA-BASED DECISION TREE TRANSFORMATION BASED ON A SOURCE SPEAKER'S SET OF HMM MODELS

7.2 Swedish HMM-based TTS

This section presents the, currently in development, Swedish HMM-based TTS (Figure 7) A complete description of the core Swedish HMM-based TTS system was adapted from previous project for Hungarian language and can be found in [10].

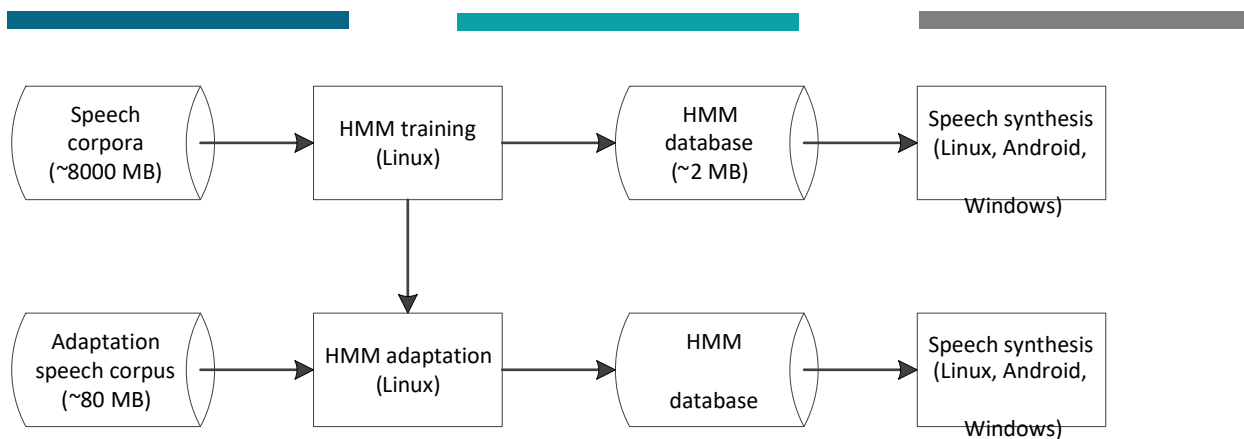


FIGURE 7 - ARCHITECTURE OF SWEDISH HMM-BASED SPEECH SYNTHESIS

The training stage is running on the Linux operating system. The synthesis stage, including the Swedish phonetic transcriber and full context labeller was running on Linux and Android 2.x platforms. As a result of the OLA project the synthesis stage is able to run on the Microsoft Windows operating system now. For this purpose, the interface and audio playback parts had to be tailored to the specifications of the given platform. BME-TMIT is currently working on the Android 4.x and Windows TTS interface for HMM-based speech synthesis.

Several language specific steps are necessary to create a Swedish HMM-based text-to-speech engine. The basics are described in [11]. In this document the most significant issues of creating an elderly voice Swedish HMM-based text-to-speech system are researched and described.

7.2.1 Speech databases

Several speech databases will be used for the average voice. Taking into account previous work in this area we expect to achieve well designed utterances and phonetically balanced sentences. The content of the utterances will be manually verified. Phoneme boundaries will be determined by forced alignment with a wide beam.

For this processing the speech databases are resampled at a rate of 16kHz on 16-bits and windowed by a 25ms Hanning-window with 5ms shift. The feature vectors consist of 39 mel-cepstral coefficients (including the 0th coefficient), logF₀, aperiodicity measures, and their dynamic and acceleration coefficients. As part of the project various voices were recorded. We defined the following requirements for HMM-TTS adaptation purposes:

- Minimum 15 minute recordings are required per speaker.
- The recordings should contain clear voice (e.g. minimize breathy and glottalized phonation).

- Background noise (including other speakers) is a drawback.
- Hesitations, repetitions and deletions should be minimal.
- At the beginning of the recordings the pauses need to be cut.

We will performed speaker adaptation based on such recordings.

7.2.2 Context dependent labelling and decision trees

Table 3 shows the context dependent labels, which were used in the design of the Swedish HMM-based TTS system. The questions for the decision tree building algorithm were defined according to these features. Depending on the parameter to be modelled (spectral, pitch, duration) the most significant question varies, although generally the questions related to phonemes are dominant. These questions are determined by the behaviour of the Swedish phonemes [12]. Table 4 shows some important features, that are used for the decision trees.

Sounds	<ul style="list-style-type: none"> • The two previous and the two following sounds/phonemes (quintphones). Pauses are also marked.
Syllables	<ul style="list-style-type: none"> • Mark if the actual / previous / next syllable is accented. • The number of phonemes in the current / previous / next syllable. • The number of syllables from / to the previous / next accented syllable. • The vowel of the current syllable.
Word	<ul style="list-style-type: none"> • The number of syllables in the current / previous / next word. • The position of the current word in the current phrase (forward and backward).
Phrase	<ul style="list-style-type: none"> • The number of syllables in the current / previous / next phrase. • The position of the current phrase in the sentence (forward and backward).
Sentence	<ul style="list-style-type: none"> • The number of syllables in the current sentence. • The number of words in the current sentence. • The number of phrases in the current sentence.

TABLE 3 - THE PROSODIC FEATURES USED FOR CONTEXT DEPENDENT LABELLING.

Phonemes	<ul style="list-style-type: none"> • Vowel / Consonant. • Short / long. • Stop / fricative / affricative / liquid / nasal. • Front / central / back vowel. • High / medium / low vowel.
----------	--

	<ul style="list-style-type: none"> • Rounded / unrounded vowel.
Syllable	<ul style="list-style-type: none"> • Stressed / not stressed. • Numeric parameters (see Table 2.).
Word	<ul style="list-style-type: none"> • Numeric parameters (see Table 2.).
Phrase	<ul style="list-style-type: none"> • Numeric parameters (see Table 2.).
Sentence	<ul style="list-style-type: none"> • Numeric parameters (see Table 2.).

TABLE 4 - THE MOST IMPORTANT FEATURES USED FOR BUILDING THE DECISION TREE.

7.2.3 Novel excitation model for Swedish HMM-TTS

Several other excitation models have also been proposed to synthesize speech with improved quality compared to the pulse-noise excitation and to lessen the computational requirements of mixed excitation. For example, Cabral uses the Liljencrants-Fant (LF) acoustic model of the glottal source derivative [13] to construct the excitation signal [13]. A strong argument for using the LF model is that the LF waveform has a decaying spectrum at higher frequencies, which is more similar to the real glottal source excitation signal [13] than pulse or mixed excitation. Drugman was one of the first researchers to create a CELP (Codebook Excited Linear Prediction) like excitation synthesis solution [14].

During analysis of speech, a codebook of pitch-synchronous residual frames (excitations) is constructed. The codebook is applied in HMM-based speech synthesis: Principal Component Analysis is used for data compression and the resulting 'eigenresiduals' are resampled to the suitable pitch and overlap-and-added together. The extended version of this method (Deterministic plus Stochastic Model, DSM) has been found to be of similar quality to a mixed excitation vocoder, while only requiring F0 to parameterize the excitation signal [15].

Ratio and his colleagues use glottal inverse filtering within HMM-based speech synthesis for generating natural sounding synthetic speech [16, 17]. Glottal flow pulses are extracted from real speech via Iterative Adaptive Inverse Filtering (IAIF, [18]), and these are used as voice source. [19] and [20] extend this model with a glottal source pulse library. Here, a library of glottal source pulses is extracted from the estimated voice source signal and used during synthesis resulting in synthesized speech having clearly better quality than traditional excitation methods. In [21], we have proposed a novel codebook-based excitation model which is based on linear prediction residual analysis and synthesis. The method has been improved and integrated to HTS in [22] which will be introduced here in detail.

7.2.4 Codebook-based excitation model

In our approach, we created a residual codebook-based excitation model that uses unit selection. Figure 8 shows the details of the analysis (speech encoding) part of the model. 16 kHz, 16-bit speech stored in a waveform is the input of the method.

First, the fundamental frequency (F0) parameters are calculated by the publicly available Snack ESPS pitch tracker with 25ms frame size and 5ms frame shift [23]. After that, Mel-Generalized Cepstrum (MGC) analysis is performed on the same frames with SPTK [24]. MGC is used here similarly as in HTS, as these features capture the spectral envelope efficiently. For the MGC parameters, we use $\alpha = 0.42$ and $\gamma = -1/3$ instead of the default HTS parameters. The residual signal (excitation) is obtained by inverse filtering with a MGLSA (Mel-Generalized Log Spectral Approximation) digital filter with SPTK. Next, the SEDREAMS Glottal Closure Instant (GCI) detection algorithm is used to find the glottal period boundaries (GCI locations) in the voiced parts of the speech signal.

The further analysis steps are completed on the excitation signal with the same frame shift values. First, a voiced / unvoiced decision is done. For measuring the parameters in the voiced parts, pitch synchronous, two period long frames are used according to the GCI locations and they are Hanning-windowed. In the unvoiced parts, a fix 25ms frame length is used. First, the gain (energy) of the frame is measured. If the frame is unvoiced, we do not apply further processing. If the frame is voiced, a codebook is built from pitch-synchronous excitation frames. Several parameters of these frames are used to fully describe the speech excitation:

- F0: fundamental frequency of the frame
- gain: energy of the frame
- rt0 peak indices: the locations of prominent values (peaks or valleys) in the windowed frame
- HNR: Harmonic-To-Noise ratio of the frame

For each voiced frame, one codebook element is saved with the given parameters and the windowed signal is also stored. These parameters will be used for target cost calculations during synthesis. In order to collect similar codebook elements, the RMSE (Root Mean Squared Error) distance is calculated between the pitch normalized versions of the codebook elements. The normalization is performed by resampling the codebook element to 40 samples. This distance will be used as concatenation cost during encoding.

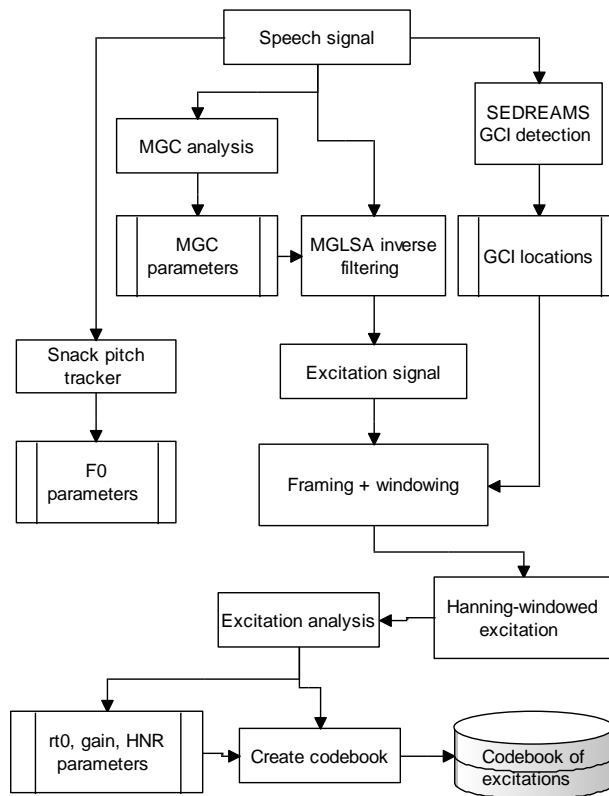


FIGURE 8 - ENCODING OF THE SPEECH SIGNAL.

Figure 9 shows the steps of the synthesis (speech decoding) stage. The input parameters are obtained during encoding (F0, gain, rt0 indices HNR) and include the codebook of pitch-synchronous excitations, too. For each parameter set, a 25ms frame is built with 5ms shift.

If the frame is unvoiced, random noise is generated with the gain as energy. If the frame is voiced, a suitable codebook element with the target F0, rt0 and HNR is searched from the codebook. We apply target cost and concatenation cost with hand-crafted weights, similarly to unit selection speech synthesis. The target cost is the squared difference among the parameters (F0, rt0 and HNR) of the current frame and the parameters of those elements in the codebook. The concatenation cost shows the similarity of codebook elements to each other and it is calculated as the RMSE difference of the pitch normalized frames. When a suitable codebook element is found, its fundamental period is set to the target F0 by either zero padding or deletion.

Next, the excitation is created by pitch synchronously overlap-adding the Hanning-windowed excitation periods. Finally, the energy of the frame is set using the gain parameter in both voiced and unvoiced regions.

The whole excitation signal is built by concatenating the pitch synchronous frames and white noise parts. Synthesized speech is obtained from the excitation signal with MGC-based filtering using the MGLSA digital filter.

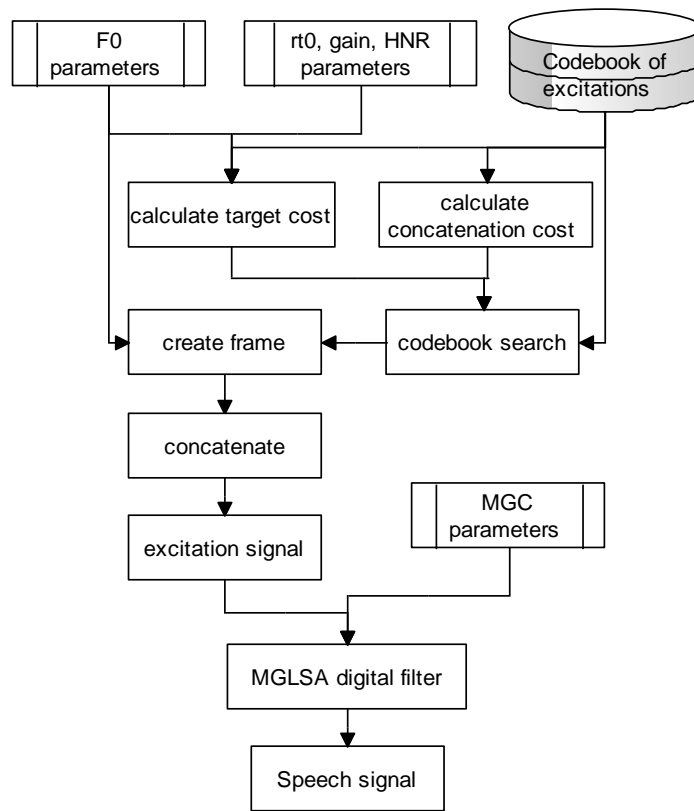


FIGURE 9 - DECODING OF THE SPEECH SIGNAL.

This model has been found to improve speech synthesis quality compared to the pulse-noise excitation [11] and has lower computational requirements than mixed excitation.



7.3 Final version of the TTS voices in Swedish, Polish and Portuguese

The final voice for Portuguese and Polish have been trained using the SPS procedure (see section 7.1).

The Swedish HMM-TTS voice is planned to be created by speaker adaptation. The average voice model was trained by the speakers and the framework that was described in Chapter 4.4. Based on the voice talent selection recordings the adaptation was carried out with the two elderly target speakers, cca. 2000 phonetically balanced sentences / speaker. The recordings were resampled at 16 kHz, 16 bits, and the phonetic labels were checked manually.

8 SDK and language packs

As the final result of the work done adapting the speech technologies for the elderly, several new Windows language packs are planned covering Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) services.

- ASR engines optimized for the Elderly:
 - Polish (pl-PL).
 - Swedish (Sw-SW). (Currently in development)
 - Portuguese (pt-PT).
- TTS voices:
 - Polish (pl-PL) male elderly voice.
 - Swedish (sw-SW) female elderly voice. (Currently in development)
 - Portuguese (pt-PT):gp
 - Two young adult male and two young female voices.
 - Two elderly male and two elderly female voices.

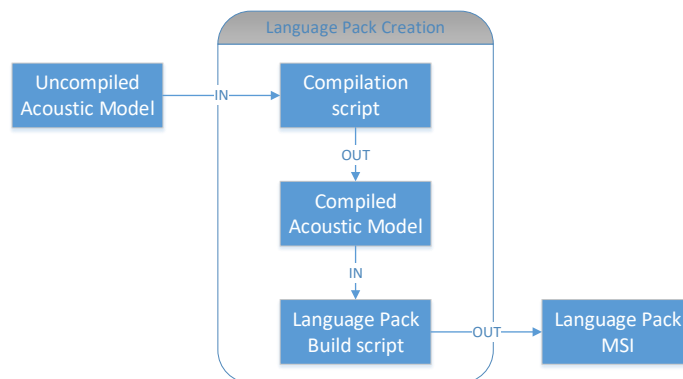


FIGURE 7 – LANGUAGE PACK CREATION

All the developed language packs are compatible with Microsoft Speech Platform V11 and are easily installed through executable MSI files. Once installed, the developers can use the language packs by following the standard guidelines from the Microsoft Speech Platform Software Development Kit (SDK).



9 Conclusion

This document tackled the development of Speech output and input interfaces, to ease the navigation in the OLA App by elder users. Two systems were envisioned an Automatic Speech Recognition (ASR) and a Text-To-Speech Synthesis (TTS) in the target market languages, Portuguese (pt-PT), and Swedish (sw-SW). For the pt-PT the development was made on especially adapted systems for the elder user, since ASR and TTS language packages are already available by technological companies such as Microsoft. As for sw-SW the focus was on the development of a new language package since this language is not currently available.



List of Figures

Figure 1 - Yourspeech architecture 12

Figure 2 - Recording of prompts using Yourspeech..... 13

Figure 3 - Synthetic voice generation..... 19

Figure 4 - Voice font building..... 20

Figure 5 - Architecture of SPS-based voice font training and speech synthesis..... 22

Figure 6 - VA-based decision tree transformation based on a source speaker's set of HMM models..... 23

Figure 7 – Language Pack creation..... 31

References

1. F. Seide, G. Li and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," in InterSpeech 2011, Florence, Italy, 2011.
2. F. Seide, G. Li, X. Chen and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in ASRU 2011, Hawaii, 2011.
3. G. E. Dahl, D. Yu, L. Deng and A. Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," IEEE Transactions on Audio, Speech, and Language Processing - Special Issue on Deep Learning for Speech and Language Processing, vol. 20, no. 1, pp. 33-42, January 2012.
4. G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, November 2012.
5. D. Braga, P. Silva, M. Ribeiro, M. Henriques and M. Dias, "HMM-based Brazilian Portuguese TTS," in PROPOR 2008, Special Session: Applications of Portuguese Speech and Language Technologies, Curia, Portugal, 2008.
6. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis," Systems and Computers in Japan, vol. 36, no. 12, p. 43-50, September 2005.
7. J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata and J. Isogai, "Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm," IEEE Audio, Speech, and Language Processing, vol. 17, no. 1, January 2009.
8. F. Z. Zheng, W. Song, F. Yu and W. Zheng, "The distance measure for line spectrum pairs applied to speech recognition," Journal of Computer Processing of Oriental Languages, vol. 11, pp. 221-225, 2000.
9. Y. Wu and R. Wang, "Minimum generation error training for HMM-based speech synthesis," in Proc. of ICASSP 2006: 89-92, 2006.
10. B. Tóth, Hidden Markov Model based Text-To-Speech Synthesis, Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics, 2013, PhD Thesis, p. 94.
11. B. Tóth, G. Németh, Hidden Markov model based speech synthesis system in Hungarian, Infocommunications Journal, Volume LXIII, no. 2008/7, 2008, pp. 30-34.
12. M. Gósy, Phonetics, The Science of Speech (in Hungarian Fonetika, a beszéd tudománya), Budapest, Osiris, 2004, p. 350.
13. J.P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In Proc. of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, 2007.
14. T. Drugman, A. Moinet, T. Dutoit and G. Wilfart, Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis. In Acoustics, Speech and Signal Processing. ICASSP 2009, pp. 3793-3796, 2009.

15. T. Drugman, G. Wilfart and T. Dutoit, A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In Interspeech 2009, ISCA, pp. 1779-1782, 2009.
16. T. Raitio, A. Suni, H. Pulakka, M. Vainio and P. Alku, HMM-Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering, Interspeech 2008, pp. 1881-1884, 2008.
17. T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio and P. Alku, HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. IEEE Transactions on Audio, Speech & Language Processing 19(1): 153-165, 2011.
18. P. Alku, Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, Speech Commun., vol. 11, no. 2-3, pp. 109-118, 1992.
19. T. Raitio, A. Suni, H. Pulakka, M. Vainio and P. Alku, Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. ICASSP 2011: 4564-4567.
20. A. Suni, T. Raitio, M. Vainio and P. Alku, The GlottHMM Entry for Blizzard Challenge 2011: Utilizing Source Unit Selection in HMM-Based Speech Synthesis for Improved Excitation Generation, Blizzard Challenge 2011 workshop.
21. T.G. Csapó, G. Németh, A novel codebook-based excitation model for use in speech synthesis, CogInfoCom 2012, Kosice, Slovakia, pp. 661-665.
22. T.G. Csapó, G. Németh, Statistical parametric speech synthesis with a novel codebook-based excitation model, Intelligent Decision Technologies, Vol. 8., No. 4., pp. 289-299, 2014.
23. The Snack Sound Toolkit [Computer program], <http://www.speech.kth.se/snack/>, accessed Sep 15, 2012.
24. SPTK working group, 2011. Reference Manual for Speech Signal Processing Toolkit Ver. 3.5, December 25, 2011.
25. Teixeira, A., Hämmäläinen, A., Avelar, J., Almeida, N., Németh, G., Fegyó, T., Zainkó, C., Csapó, T., Tóth, B., Oliveira, A., Sales Dias, M. (2013) Speech-Centric Multimodal Interaction for Easy-to-Access Online Services – A Personal Life Assistant for the Elderly, in Proceedings of the International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion (DSAI), Vigo, Spain.
26. Annika Hämmäläinen, Jairo Avelar, Silvia Rodrigues, Miguel Sales Dias, Artur Kolesiński, Tibor Fegyó, Géza Németh, Petra Csobánka, Karine Lan and David Hewson, "The EASR Corpora of European Portuguese, French, Hungarian and Polish Elderly Speech", LREC Conference, 2014.