Technology and Corpora for Speech to Speech Translation
http://www.tc-star.org

Project no.:	FP6-506738
Project Acronym:	TC-STAR
Project Title:	Technology and Corpora for Speech to Speech Translation
Instrument:	Integrated Project
Thematic Priority:	IST

Deliverable no.: D9
Title: TTS Progress report

Due date of the deliverable:	31st of March 2005
Actual submission date:	17 May 2005
Start date of the project:	1st of April 2004
Duration:	36 months
Lead contractor for this deliverable:	UPC
Author(s):	Antonio Bonafonte, Harald Höge, Imre Kiss, Asuncion Moreno, David Sündermann, Ute Ziegenhain, Jordi Adell, Pablo D. Agüero, Helenca Duxans, Daniel Erro, Jani Nurminen, Javier Pérez, Guntram Strecha, Martí Umbert, Xia S wang

Revision [final version]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Table of Contents

1 Overview

The main objective of this work package is to produce synthetic speech with improved quality and added functionalities for speech-to-speech translation. Speech-to-speech translation is a scenario that makes speech synthesis harder because TTS systems are designed for grammatical and "correct" inputs. In speech to speech translation the input is not well formed because it comes from transcription of speech, not text. Furthermore, speech recognition and spoken translation engines make errors. As a consequence, the input is far from being grammatically correct, which is the situation where most of the systems, are trained. On the other hand, the speech-to-speech translation framework offers a valuable source of information: the original voice. The voice characteristic is an information source which may allow generating richer prosody being able to transmit not only linguistic but paralinguistic information (as the attitude of the speaker) which is crucial for successful communication. Furthermore, in TC-STAR we want to analyse information in the source voice for transforming the synthetic voice so that it sounds like the original speaker, creating a new functionality that allows voice customization of speech synthesis. This capability may turn crucial when several people use the translation system as is the case of meetings or parliamentary debates.

Coordination

To coordinate the work in the work package, several meetings have been arranged. In the general meetings of the project (May in Trento and November in Barcelona) parallel meetings have been allocated. Furthermore three specific meetings for the work package have been organized: Maribor, July 2004; Barcelona, December 2004; Dresden, Mars 2005. A distribution list has been set up from the beginning of the project (wp3@tc-star.org) and a web site (internal part of www.tc-star.org) . Several conference calls have been arranged with participants of the participants on the workpackage.

Mobility

For exchanging know how in the field of voice conversion and manipulation of speech segments, the PHD-Student David Sündermann stayed for several months on the sides: UPC; Siemens and RWTH. The success of the close cooperation can be seen on the common papers made by researchers from those sides (see references, chapter 7).

Subcontracting

For performing some of the task involved in language resources (WP4) Siemens and UPC plan to subcontract external subcontractors.

- Siemens: Recording for synthesis voices, creation of the corpus 'frequent used sentences', pitch annotation, lexicon
- UPC: labeling of speech data

Collaboration with other organizations

The partners have founded an external consortium (European Center of Excellence for Speech Synthesis – ECCES- : www.ecess.org), open to any research group, and with more than 10 research groups involved. The goal of this consortium is to increase the critical mass in speech synthesis.

The consortium has agreed to define a modular system with clearly defined interfaces. In this way, modules from one system can interact with modules from other system to build the best final system. This allows to institutions with small critical mass to focus in their specific research topic and participate in evaluation campaigns without the overhead of develop and maintain a complete state-of-the-art system. This consortium has actively participated in the discussions about the specifications on language resources, evaluation and architecture design. In fact, all the partners plan to participate in evaluation campaigns organised by TC-STAR and some partners plan to produce databases in their national language using the specifications from TC-STAR.

Review of the tasks in the Work package and structure of the document

The first task of the workpackage "*Specification on language resources and evaluation*" has been completed.

The specifications on Language resources define all the steps needed to produce high quality data for generating high quality synthetic voices (in the framework of unit selection) including corpus design, speaker selection, recording platform, speech annotation, lexicon and the procedure to validate the produced resources. The first part of the specifications aims to create baseline voices. The second part address specific research activities of the project: voice conversion and expressive speech. Section 2 of this document summarizes the specifications and also the progress done in producing the language resources. The full version of the specifications can be found in deliverable D8.

With respect to the specifications to evaluate speech synthesis, several tests have been defined to evaluate the speech synthesis component (black box), the speech synthesis modules (glass box) and two specific research activities: voice conversion and expressive speech. Section 3 goes over the main points of the evaluation specifications. Deliverable D8 also includes the complete specifications.

Section 4 reviews the work on the second task of the work package: baseline systems for research and evaluation. The TTS system has been divided into three functional modules, text processing, prosody generation and acoustic synthesis. The interfaces (input/output) have been formally defined. This allows to evaluate each module separately (diagnosis) and to compose modules from different partners to get the best result or to combine functionalities. With respect to development of the systems, each partner has advanced on their own system in order to have TTS in the three languages of the project (English, Mandarin and Spanish). In state of the art concatenative systems using unit selection, the performance of the system depends on the speech voice: the baseline systems will be operative and evaluated once the language resources are finished (already started).

The workpackage devotes special effort to two research activities. Section 5 and 6 are devoted to the progress on these activities. In particular section 5 deals with the task *Voice conversion, manipulation and compression.* The partners have investigated in the speech representation and the conversion of the parameters. The final goal in TC-STAR is to achieve cross-language voice conversion so that the output synthetic speech sounds like the original natural speech. However, in this first year all the voice conversion activity has been focused in intra-lingual voice conversion. Nevertheless, some effort has been devoted to text-independent voice conversion. Here, the training date required to estimate the conversion function does not need to correspond to the same text. This can be seen as a first step towards cross-lingual voice conversion. The baseline algorithms are been set up for the three partners. In general the systems successfully transform the identity of the source speaker into the target one. However, the quality of the voice is degraded significantly with respect to natural speech. Alternatively, some research has been done that gets good quality but the voice of the transformed speech resembles a third speaker. Additional work

has to be done to accomplish both goals, quality and identity change, using the same technique. Section 5 review the contribution in several aspects as transformation of LSF parameters, transformation or prediction of the target residual signal, voice conversion using vocal tract normalization (VTN). Also some promising results have been obtained in compression. The goal is to produce low footprint systems which can be used in embedded systems. The work is based on selection of speech units and coding using adaptive multirate. Analysis algorithms have been developed to represent speech using the glottal-filter approach and the deterministic + stochastic model. It is expected that this representations allow better manipulation with respect to prosodic changes, concatenation and voice conversion.

Section 6 is devoted to the last task of the workpackage, *prosody modelling an expressive speech*. The information of speech voice is not limited to pure linguistic but includes other information as the attitude of the speaker towards the message or towards the audience, the feelings of the speaker, etc. It is difficult to get this information from the text: first the semantic and pragmatic analysis is still far from been a mature science. Even more: in many cases this information is not present in the text. Transcriptions of speeches can indeed be misunderstood because this information is not coded in the text. In order to improve the expressivity of the synthetic speech we have defined a framework to include information from the source speech. The idea is that some acoustic features in the source speech can be used to derive acoustic features in the target speech (in different language). To implement this idea it is required to analyse the source speech, to map the information from source speech to target text and to generate the prosody. During the first year most of the work has been devoted to prosody generation. A new paradigm has been proposed to that achieves better results. The idea is to integrate the analysis of the contours (feature extraction) and the generation of the model. The basic idea is to define first classes (model) and to find the best analysis for all the members of the class (feature extraction). This implies that the feature extraction is done simultaneously for the entire corpus and not sentence by sentence. This paradigm avoids the use of stylization and voiceless interpolation which sometimes are hand tuned and influence significantly on the final models. Furthermore the results are more robust to errors either in the estimation or in the linguistic features. This paradigm has been applied to several intonation models (Tilt, Fujisaki, Sup-Bezier). The investigated methods improve the synthetic speech in general (using text as input). Additionally, it is possible to extend them including additional features derived from the input speech. Furthermore, some work has been done to analyse the input speech and detect features as prominence, speech rate, and f0_contour, including coding the f0_contour using a discrete and limited alphabet.

Dissemination

The work that has been done in the work package has been disseminated to the scientific community. The partners have already published the main results of the work in scientific conferences (see references) and in workshops. Furthermore, the results have been communicated to members of ECESS and to members of the NoE Similar. This two consortiums include many laboratories in Europe and some from outside Europe.

Main Achievements

To end this overview, the following list summarizes the main achievements accomplished during the first year of TC-STAR in the work package *Speech Synthesis:*

- Infrastructure
    - Specifications of LR finished
    - Evaluation procedures finished

  - Functionalities and Interfaces between modules finished
  - Starting of production of LR

- Research on Baseline, English, Mandarin and Spanish, using existing LR

  - Modules structure compatible with defined interfaces
  - Algorithms for text processing, prosody and acoustic synthesis

- Research on voice conversion, compression and manipulation with existing LR

  - First steps towards cross lingual voice conversion: Text independent approach
  - Improving voice quality by residual prediction for voice conversion
  - Improvement to map speaker's voice characteristics: Integration of phonetic information
  - Footprint reduction and quality improvement: Integrated approach on speech coding and manipulation
  - Exploring new models on speech production for voice conversion.

- Research on prosody modelling and expressive speech

  - Preserve naturalness in prosody for tonal languages: eigen pitch approach
  - Definition of a novel joint feature extraction and modelling approach (JEMA)
  - Improve prosody generation by an application of (JEMA) to three existing models

2 Specification of Language resources

All WP3 partners (Nokia, Siemens, SPEX, UPC) contributed on the specification of the language resources. This work was coordinated and drafted by Siemens.

This chapter contains also the information about the work done for the production of LR needed for speech synthesis. This concerns tools needed to produce the LR and the status of production.

2.1 Specification of Language Resources for Speech Synthesis

Language resources are needed to build speech synthesis systems and to make research in speech synthesis. The basic set of language resources needed in synthesis is

- lexica containing entries for pronunciations with stress marks and POS tags
- voices defined by recorded and annotated speech
- tagged corpora

Within TC-STAR language resources for advanced, high quality speech synthesis systems and for making research in voice conversion and in expressive speech have to be provided for the languages UK-English, Spanish and Mandarin.

All these language resources have to be specified and produced if not available. Concerning lexica it was decided to use the specification of LC-STAR[1] and to use the existing lexica as far as suitable. The specification and production of voices is done within TC-STAR. The new specifications are part of the deliverable D8. For tagged corpora existing languages resources are used.

In the following section 2.2 a short overview of the content of specification of voices as documented in D8 is given. The TC-STAR partners decided to make a new specification of language resources for voices, which is suited for the tasks to be done. Main reasons to make a new specification were:

- Lack of precision (e.g. no requirements for recording equipment, no specification of pitch marking)
- Lack of completeness (e.g. missing specification of LR for voice conversion)
- Lack of validation criteria (minimal requirements) to assure quality

Due to the experience made in the EU-funded SpeechDat[2] projects to specify language resources for ASR it was clear, that the generation of such specification would be a hard and time consuming task. The specifications delivered should have a quality with the potential for becoming a quasi standard on which LRs in a variety of languages can be produced.

2.2 Deliverable D8; LR specification part

The part of D8 specifying language resources covers the language independent part (LIP) of the specifications of language resources for voices. Language specific issues and language specific deviations from the language independent specifications are described by each provider of a language resource in a separated document LSP (LSP denotes the Language Specific Part).

The creation of the TC-STAR voices for TTS systems and research on voice conversion is based on read speech**.** For this issue, text corpora are specified which have to be read by selected speakers. For research in expressive speech recorded data (e.g. recordings from the Spanish or European parliament) and read data will be used.

The main chapters of the specifications are:

- the construction of the text corpora to be read
- the procedure to select suited speakers
- the recording platform
- the annotation of the recordings of the speakers
- the database interchange format

In order to make available high quality language resources the specification the language resources created will be **validated**. Specific validation criteria have been developed.

[1]http://www.lc-star.com/WP2_deliverable_D2_v2.1.doc; for Mandarin and Spanish already validated LC-STAR lexica exist.

[2] www.speechdat.org

The main issue in synthesizing speech from any domain is to achieve a good coverage on speech segments with all their prosodic properties used in a given language. As speech segments, triphones or syllables in various prosodic contexts are regarded. Due to the large amount of such units and given a restricted corpus, 100% coverage is hardly to achieve. A compromise concerning effort and coverage has to be made. For the baseline of the synthesis system it was decided to record 10h of speech for each voice. This amount corresponds roughly to 90 000 spoken word tokens to be read by a baseline speaker. The text corpus consists on text derived from transcribed speech (45 000 word tokens) to cover phenomena found in speech, on written text (27 000 word tokens) and on specifically constructed text (18 000 word tokens). The last corpus is designed to achieve high coverage and for research in intra lingual voice conversion. Part of the text of the transcribed is translated in order to achieve parallel corpora needed for cross language voice conversion.

For the baseline systems for each language one male and one female speaker and for research in voice conversion 4 bilingual speakers per language pair (Spanish-UK-English; Mandarin-UK-English) are selected[3].

The usefulness of the recorded speech depends on the quality of the speech signal and on the precision with which the glottal closure can be reliable marked (pitch marking). The recordings have to be made in high quality studios with low noise level and low reverberation time. Finally the recorded speech has to be annotated. All speech has to be completely phonetically transcribed, segmented and pitch marked.

2.3 Status of production: Text Corpora, related Voices and lexicon

The status on the text corpora and the related baseline voices is as follows

Partner	**Language**	**Text Corpus**	**Lexicon**	**LSP**	**Speaker selection**	**Recordings**
Nokia	Mandarin	50% ready	in progress	draft	in progress	in progress
Siemens	UK-English	50% ready	not started	draft	in progress	in progress
UPC	Spanish	finished	in progress	draft	in progress	in progress

Specific work done for designing the text corpora is described in the following sections.

2.3.1 Nokia

Foreword

[3] Due to the funding situation for Mandarin only 1 voice is created. Within the framework of ECESS a second voice will be created. The same holds for voice conversion, where for the language pairs Mandarin-UK-English only 2 speakers are provided within TC-STAR.

Mandarin Chinese text materials face special challenges since the parliamentary speech and transcribed text resources are not easily available. Recently Recently (beginning of April) we reached an agreement with ChineseLDC about the license of their transcribed broadcasting news corpus for Nokia and other partners in TC-STAR and ECESS. However, we haven't received the corpus yet. The coverage figures presented here are mainly for C2_T and C3.3_T.

Tonal Syllable Coverage in Chinese

The aim of this section is to do a study on the Mandarin Chinese tonal syllables. The approximately 38100 Chinese most common words, taken from the LC-STAR project, have been analyzed in order to get the necessary tonal syllables to be covered within the whole corpora sentences (90000 running words), as well as their apparition probabilities. The outcome of this first study is shown in Table 2.3:

Different tonal syllables in Modern Chinese Lexicon– N1	Different tonal syllables in LC-STAR Common Word Lexicon – N2	LC-STAR Common Word Lexicon Coverage of tonal syllables
1288	1243	96.5%

Table 2.3 Tonal Syllable Coverage in Chinese LC-STAR Common Word Lexicon

The aim of the synthesis corpus will be to always achieve, at least, the 95% of the tonal syllables of LC-STAR, that is to say not less than 1181 syllables.

We have made an analysis for C2_T Novels and C3.3T Mimic sentences.

Finally, the corpora sizes (in words) that have been used and tried to reach as target are shown in Table 2.4 :

	Initial Corpus size in sentences	Initial Corpus size in words	selected corpus in words	Tonal syllables	Tonal syllable coverage for LC-STAR
C2_T			27049	1146	92.3%
C3.3_T	540465	5642073	3007	1120	90.1%

Table 2.4 : Initial and Target Corpus sizes

2.3.2 Siemens

Coverage experiments in UK-English.

In order to generate the corpus C_T for UK-English fulfilling the coverage criteria defined in [Bon05], we required two kinds of language resources: a UK-English phonetic lexicon and large text corpora of the specified domains from which the C_T corpus can be extracted. In the following two subsections, we describe the generation of the respective material and its properties. Then, we report on the triphone coverage experiments and investigations on the coverage of rare phonemes.

The Phonetic Lexicon

In order to produce a preliminary phonetic lexicon for US-English, we merged the UK-English version of the CELEX [Baa93] and the Unisyn [Fit00]; both were provided for research purposes only. This was done in the following way:

- At first, we mapped the lexicons' phoneme sets to the one that is to be used for TC-Star UK-English speech synthesis.
- After merging the lexicons, a lot of word entries were represented by several transcriptions. However, for the coverage experiments, we agreed to use only the most frequent transcription of each entry, hence, the others were deleted.
- To obtain a phonetic lexicon with entries according to the LC-Star paradigm, we applied the word list of the LC-Star US-English phonetic lexicon (50,466 words) to the merged UK-Lexicon resulting in a sub-lexicon containing 46,489 words. This lexicon is to be used as standard phonetic lexicon when performing coverage experiments and generating the C_T corpus.

The Text Databases

The C_T corpus is to be derived from three domains: parliamentary speeches, novels, and frequently used phrases. As the latter are manually generated, we are only able to utilize the first two domains, where we are provided sufficient text material for selecting certain sentences according to the required coverage. For the parliamentary speeches, we used the UK-English part of the EPPS corpus, version 2005-02-24 [Gol05], 30,366,390 running words. As novels, we used the collected works of A. C. Doyle, 723,552 running words.

Experiments on triphone coverage

At first, we transcribed the whole EPPS corpus by means of a simple lookup into the above described standard lexicon; in doing so, we ignored unknown words. Now, we checked how many different triphones from the lexicon are contained in certain subsets of the transcribed text. The number of running words in the considered subsets were logarithmically varied between 1,000 and the maximum possible (the entire EPPS). In Table 1, the results are shown with and without stress. In the lexicon, we have 12,689 different triphones (or 20,888, when taking the stress into account).

Looking at the highest seen triphone coverage (83.3 %), we note that the finally required 90% could be achieved by exploiting the triphone coverage corpus C3.2_T selecting (90%-83.3%)/100%*12,689triphones = 850 triphones not yet covered by the corpus from the lexicon, i.e., at the most 850 words. These words have to be contained in the C3.2_T (whose size is about 8000 words) that has to be generated using very large databases as the internet.

Besides, it should be mentioned, that, in the future, only non-singleton lexicon triphones are considered resulting in an essential relaxation of the coverage problem.

running words of the EPPS	different triphones (with stress)	coverage (with stress) / %
1k	780 (834)	6.2 (4.0)
10k	2,407 (2,842)	19.0 (13.6)

100k	4,538 (5,950)	35.8 (28.5)
1M	7,247 (10,606)	57.1 (50.8)
10M	9,634 (14,852)	75.9 (71.1)
30M	10,572 (16,741)	83.3 (80.2)

Table 2.2 Triphone coverage for UK-English, with and without stress.

Experiments on the coverage of rare phonemes

When generating the corpus C3.3_T, we had to make sure that each phoneme occurred at least 10 times, cf. [Bon05]. This was to achieve a minimum coverage of rare phonemes. To fulfill this criterion, we performed the following steps:

- From the transcribed EPPS corpus, we selected the paragraphs with 10 to 20 words obtaining a subcorpus of 1,360,773 words.
- Now, we applied a greedy algorithm that, in each iteration, selected that sentence whose transcription maximized the ratio between the number of the currently rarest phoneme according to the already extracted sentences in the instantaneous sentence and the number of words thereof. The algorithm iterated until the number of extracted words was 2008.

The rarest phoneme of the C3.3_T occurs 62 times, so that the coverage criterion is fulfilled.

2.3.3 UPC

Triphone coverage in Spanish

The aim of this section is to do a study on the Spanish common triphones and diphones and to have a reference list. The approximately 50000 Spanish most common words, taken from the LC-STAR project (with a proven coverage on several texts greater that 95%), have been analyzed in order to get the necessary triphones to be covered within the whole corpora sentences (90000 running words), as well as their apparition probabilities. The outcome of this first study is: Different triphones: 7951, Singletons: 965, Triphones (without singletons): 6986

The aim of the synthesis corpus will be to always achieve, at least, the 95% of the triphones (without singletons), that is to say not less than 6637 triphones.

To prove the suitability of these triphones, a coverage test has been done with 2.735.000 words taken from parliament texts. The phonetic transcription has been obtained by using the Saga software. The results are: Different triphones: 8759, Singletons: 508, Triphones (without singletons): 8251

The results from the LC-STAR analysis and the parliamentary texts were compared in order to know how many triphones were in both lists. There are 6551 triphones that appear in both lists of triphones (without singletons). These have text coverage of the 98.79% of the triphones, which overcomes the minimum desired percentage of the 95%. On the other hand, 1700 triphones from the text did not belong to the common triphones list, and 435 needed triphones to cover did not appear in the analyzed text.

Prosodic coverage

Prosodic coverage for Spanish is done based on diphones and their position in a sentence. For this purpose we define:

- Voiced diphones: one phoneme is voiced
- Unvoiced diphones: Both diphones are unvoiced
- Position of diphones in sentences:
  - Initial : From the beginning of the sentence till the first stressed diphone (included)
  - Prepausal: From the last stressed diphone till the end of the sentence
  - Middle: from the first and last stressed diphones (not included)

In order to define a significant diphone set a selection of distinct diphones (stress and unstressed with a frequency of occurrence ($fd>10^{-4}$) in the parliamentary texts was done.

The final corpus should accomplish to find 2 examples of the following list: Unvoiced no-prepausal, Unvoiced prepausal, Voiced initial, Voiced middle, and Voiced pre-pausal

As a result, a text of approximately 1.700.000 words has been analyzed in order to get the reference list of diphones needed to cover. A total amount of 451 diphones (433 voiced and 18 unvoiced) has been obtained, which become 2201 diphones when taking into account all possible positions in a sentence.

Selection procedure

The aim here is to choose several sentences from different text sources in order to cover the desired triphones and diphones, including interrogative sentences. Each separate corpus has been mainly obtained by using an in-house corpus balancing tool, CorpusCrt.

C1.1_T corpora : Parallel Transcribed Text

The procedure to obtain C1.1_T has been simply to apply the corpus balancing tool to the input file, (about 2.700.000 words) having previously ordered the sentences inversely from the triphones probabilities point of view, so that the phonetically less probable sentences were at the top of the input file. This strategy ensures that CorpusCrt will take from the beginning the less frequent triphones.

C1.2_T corpora : General Transcribed Text

The procedure to obtain C1.2_T has been to divide it (about 8.800.000 words) in 2 subcorpus. The first one (C1.2a_T) is aimed to cover as much missing triphones as possible, with some sentences that contain question marks. The second one (C1.2b_T) is mainly focused on covering the missing diphones on positions with at least 2 apparitions. The first one represents around the 70% of the whole C1.2_T and the second one approximately the 30%.

C2_T corpora : Novels with short sentences

Finally, the C2_T is obtained similarly than before. First of all, those sentences that do not introduce any new triphone or diphone are erased. This makes the initial corpora diminish from 315.000 to 260.000 words. Then, CorpusCrt is applied.

C3_T corpora : Selected Sentences and frequent phrases

Those missing triphones and diphones to cover are now covered with selected sentences written on purpose.

Results

The obtained corpora in number of words and sentences is summed up in Table 2.1 :

	# sentences	# words	# different triphones
C1.1_T	213	9419	3768
C1.2a_T	288	25257	1248
C1.2b_T	147	10849	253
C2_T	545	27030	1047

Table 2.1 : Number of sentences, words and additional triphones for each corpus

Triphones coverage

The target number of distinct triphones is 6637. In the C1_T corpora, 5269 triphones have been covered (which represent the 75,4% of the target). These triphones cover the selected sentences at a 98.71%. With the C2_T corpora, 1047 triphones more are covered, having now a total amount of 6316 (90.41% of the target) . These triphones cover the selected sentences at the 98.49%.

Interrogatives sentences

In the C1.2a_T corpus, 66 out of 288 sentence contain question marks, which represent the 21% of the sentences.

Position on diphones

The target number of diphones in different positions was 2201 and a 67.92% was achieved wirh C1 and C2

C3_T corpora

There were still 706 diphones and 670 triphones missing in the previous corpora that are covered with selected phrases. Sentences are manually selected (8000 running words)

A number of frequent phrases are built in the following domains: Brands, Foreign countries and major cities, Spanish provinces and cities in several prosodic positions, Digits and Cardinals, Ordinals with common Spanish proper names (gender is included),

Spellings, Dates, Questions, WEB, email and URL addresses, and Dialogue corpus.

2.4. Tools

Several tools have been developed. Main cost factor on production is segmentation on phoneme level and epoch detection. Several segmentation algorithms were investigated. To study the quality of epoch detectors manually marked databases are started to be produced.

2.4.1 UPC

Automatic phone segmentation for TTS synthesis.

When using concatenative TTS synthesis, we need to spend a big part of the effort on preparing the database. Parts of this process use to be completely manual or manually supervised. Phone segmentation is one of the tasks that require largest effort .

In our work we have compared three classical methods and a proposed one and evaluated them objectively [Ade04]. Then another new method is proposed and together with previous ones they were both objective and perceptually evaluated [Ade05]. The results show that the quality of the segmentation needs to be evaluated not only using objective evaluations but also subjective ones.

Baseline methods studied:

Hidden Markov Models: It consists on performing a forced alignment by means of the Viterbi algorithm. It is assumed that the phonetic transcription is known, Transitions between models are then considered as phone boundaries. [Tay91]

Dynamic Time Warping : This method uses a dynamic time warping algorithm to align synthesized voice with a non-segmented one. In TTS the database is labeled so we know where the phone boundaries are in the synthesized speech. Then, these boundaries are mapped onto the recorded speech by means of the alignment performed. [Kom03].

Artificial Neural Networks: They can be used to correct the boundaries given by HMM based systems and achieve better performance .ANN try are designed to estimate the probability of having a boundary in a specific frame from a set of acoustic characteristics extracted from the voice. HMM boundaries are moved to the closest maximum given by the network [Tol 03].

Proposed methods:

Acoustic Clustering-Dynamic Time Warping (AC-DTW): Phonetic boundaries are established by a Dynamic Time Warping algorithm that uses the *a posteriori* probabilities of each phonetic unit given an acoustic frame. These *a posteriori* probabilities are calculated by combining probabilities of acoustic classes, which are obtained from a clustering procedure on the acoustic feature space, and the conditional probabilities of each acoustic class with respect to each phonetic unit [Góm02].

Regression Tree-Boundary Specific Correction (RT-BSC): HMM boundaries are refined using phonetic features (i.e. manner, articulation point, voice, etc. . .) of both phones involved in the transition. A small sub-corpus is used to train a Regression Tree that makes a regression of the error between the manually supervised and the HMM-based segmentation as a function of the phonetic features by means of binary questions. Then, this tree can predict the error for the rest of the corpus, thus it can be corrected.

A further comparison between ANN and RT-BSC methods pointed out that phonetic features are better suited for HMM boundaries refinement than acoustic ones [Ade04]. Our proposed methods have overcomed classic methods performance on objective evaluation. However, a perceptual test showed that none of the methods improved the overall quality of the system. [Ade05]

3. Evaluation

In deliverable D8 we have defined several tests to evaluate speech synthesis in the context of TC-STAR. The development of speech technology in TC-STAR is evaluation driven. Assessment of speech synthesis is needed to determine how well a system or technique compares to others or how it compares with previous version of the system.

In order to make a useful diagnose of the system in TC-STAR we will not only make a test of the whole component but also specific tests for each module of the speech synthesis system. In this way we can assess better the progress on specific modules. Furthermore, it allows identifying the best techniques in the different processes that are involved in speech synthesis. To allow the comparison of different modules we have defined a common specification of the modules and specific test for their evaluation.

Text-to-speech systems perform a range of processes, from text normalization, pronunciation, several aspects on symbolic and acoustic prosody, etc. Finally we are interested on the *quality* of the overall system. However, the evaluation of the whole (*black box evaluation*) does not allow pinpointing which part of the system causes the most relevant problem. Furthermore, this method does not allow participating on the evaluation to small teams of researchers whose specialty of research is in one specific topic. In TC-STAR we will certainly evaluate whole systems, but we also want to evaluate different tasks to drive more valid conclusions about the results of different algorithms. Defining *modules,* with well defined input and output allows keeping constant all the modules except one and comparing the results caused by the algorithms involved on that module (*glass box* evaluation).

There are many processes involved in speech synthesis. Researchers working in a particular one would prefer to make specific tests to evaluate that process. For instance, some tests have been proposed to evaluate each aspect of prosody, from intonation, pausing, accentuation, etc. However, from a pragmatic point of view, when designing a general evaluation framework, the number of modules needs to be limited. The evaluation of speech synthesis involves in many cases human evaluation and is needed to limit the number of test for each campaign. Also, in order to compare different systems only generic modules can be defined because not all the systems are built up of the same processes. Furthermore, although we assume that speech synthesis is built up of independent modules, in fact this is not absolutely true. For instance, a promising area of research is modeling the correlation between the different features related with prosody (f0, duration, etc.). Keeping these processes together allows modeling this interaction. Therefore, there is a compromise in the number of modules. In TC-STAR we define three broad modules: symbolic preprocessing, prosody generation and acoustic synthesis. The modules have been defined through their interfaces, i.e., the formal description of the input and output.

Finally, in WP3, two specific areas of research are voice conversion and expressive speech synthesis. In order to evaluate the research specific tests have been defined.

The following tests have been defined (cf. D8):

3.1 Evaluation of modules.

Module 1: Text analysis. The goal of the text analysis is to transform the orthographic input string to the representation of the sounds. It involves text normalization, which transform ambiguous text such as numbers, dots and abbreviations into non-ambiguous words (which are known as "standard words"). In the case of Mandarin, this module segments the character stream into words. This module also copes with grapheme to phoneme conversion and with the assignment of lexical stress and syllable boundaries. Furthermore, this module tags the words with the POS (part-of-speech label), which is needed for prosody assignment.

- Test M1.1: Text Normalization
- Test M1.2: Word Segmentation (Mandarin)
- Test M1.3: Evaluation of POS-tagger
- Test M1.4: Evaluation of grapheme-to-phoneme

Module 2: Prosody. The output of the second module is *acoustic prosody* (cf. D8), defined as the F0-contour, intensity contour and phoneme (+ pause) duration. Other parameters as voice quality can be included in the second phase of the project.

- Test M2.1: Evaluation of prosody (using segmental information, resynthesis)
- Test M2.2: Judgment test using delexicalized utterances
- Test M2.3: Functional test using delexicalized utterances (identify written sentences which the produced delexicalized prosody)

Module 3: Speech generation. The third module produces speech from the phonetic and (acoustic) prosody description. Segmental quality or segmental identification is one of the main factors in getting good overall quality. Intelligibility and quality are the two needed characteristics of the produced voice.

- Test M3.1: Evaluation of speech generation module: functional test (transcribe semantically unpredictable sentences).
- Test M3.2: Evaluation of speech generation module: judgment test (naturalness and intelligibility)

3.2 Evaluation of specific research topics.

As mentioned before, specific test have been defined to evaluate some specific research activities in the project.

Voice Conversion. Voice conversion is the adaptation of the characteristics of a source speaker's voice to those of a target speaker. The final goal of the project is to adapt the speaker characteristics using few data from the target speaker and with source and target in different languages. However, in the first evaluation the task is limited to intralingua voice conversion. The evaluation criteria is the speaker identity (goal) but also the quality of the voice.

- Test VC.1: Evaluation of research on voice conversion excluding prosody. (all the comparisons use the same prosody so that it does not influences in the identity judgment)
- Test VC.2: Evaluation of research on voice conversion including prosody. (comparisons use natural prosody: includes work in prosody adaptation)

Expressive speech. Most of the evaluation procedures in expressive speech are functional tests related with emotion: synthetic speech is produced using one of a given predefined set of emotions. The subjects are asked to identify the emotion on the speech (close set answer). The aim of TC-STAR is not to produce emotional speech but expressive speech. One characteristic of expressive speech is that it can signal para-linguistic information using prosody. Produce expressive speech from general text requires very high knowledge of the world and high cognitive capabilities. However, in TC-STAR we want to explore how some para-linguistic information can be derived from the source speech and used to produce the synthetic voice.

- Test ES: Evaluation of research on expressive speech (judgment test about the expressivity of the voice and the appropriateness of the expression given the content)

3.3 Evaluation of speech synthesis component.

The speech synthesis systems are evaluating using a test based on ITU.P85 (judgment test over several aspects of the synthetics voice)

4. Baseline systems for research and evaluation of speech synthesis

The creation of TTS baseline systems for UK-English, Spanish and Mandarin is highly dependent on the availability of language resources. The specifications for LR collection have been finalized (see D8 of WP3), but the LR creation work is still in progress in the project. However, certain algorithm development for the baseline systems can be done without the final LRs. This section describes the contribution of the WP3 partners: Nokia, Siemens and UPC for baseline system development during the first project year.

4.1 Nokia

During the reporting period, Nokia has further developed its existing waveform concatenative TTS system, which it intends to use as baseline for TTS evaluations. In legal terms, Nokia will utilize amalgamated SW for the TC-STAR evaluations. All parts of the system (text processing, prosody modeling and acoustic synthesis) have been under development. Since the production of language resources is still ongoing in the project (see Section 2), Nokia has preliminarily relied on its in-house Mandarin language components for the work. We have had four conference publications in the TTS area during the reporting period.

Work on TC-STAR architecture

The system has been brought to the modular structure as proposed for TC-STAR and ECESS. However, the TC-STAR evaluation API is not yet integrated. This work is planned for the following months.

Work on Text Processing Module

On the text pre-processing front, Nokia improved its existing framework by refining text processing rules and grapheme-to-phoneme (GTP) conversion. Some work has been spent on defining an automatic scheme for optimal selection of training data for GTP [Tia05], and on an improved syllabification method [Tia04a]. Moreover, Nokia investigated footprint reduction

techniques for TTS pronunciation dictionaries [Tia04b]. Some of the above results are applicable also to other languages than Mandarin.

Work on Prosody Module

Concerning prosody modeling, Nokia improved its basic statistical framework for prosody modeling. A special emphasis has been put on the representation of pitch information, which is especially important for Mandarin, being a tonal language.

Nokia introduced a novel concept called syllable-level eigenpitch [Tia04c]. With this new parametric representation, the tonal patterns of the language are well preserved.

Work on Acoustic Synthesis Module

The research work on acoustic synthesis has been focused on two topics in the reporting period. First, the distortion measure used in unit selection was refined to use an optimum combination of pitch, duration and context information. Second, Nokia investigated alternative methods to PSOLA-based unit concatenation and smoothing for synthesis. In addition, some effort has been spent on the refinement of in-house Mandarin resources, namely, a subset of syllable units were picked from the acoustic database to provide better quality synthesis and lower footprint.

4.2 Siemens

Siemens has developed and is further developing a speech synthesis system called 'Papageno_embedded' for several languages tuned to embedded systems. Due to the restrictions in memory and processing power the system is based on diphone synthesis, leading to restricted speech quality. Before the start of TC-STAR preliminary work has been started [Hol00] to develop a system called Papageno-Server dedicated to generate high quality speech synthesis. The R&D done within TC-STAR where Siemens is responsible for developing a speech synthesis baseline system for UK-English, is a continuation of the work done on Papageno-Server. In legal terms Siemens will provide amalgamated software to TC-STAR.

The Papageno technology, which is common for both systems, is based on the following principles:

- Separation of language independent program code and language resources as far as possible
- Using an architecture based on the modules: text processing, prosody generation and acoustic synthesis.

This approach comes close to the architecture as proposed within ECESS, which is used within TC-STAR. In the first year the work was focused on supporting the envisaged TC-STAR architecture and optimizing the text processing module.

Work on TC-STAR architecture

The basic operations of a system as loading the language resources, starting and stopping the system have been made available on the interfaces of the system modules. Due to this work the modules are well encapsulated with well-defined interfaces leading to modules, which can be easily exchanged as foreseen in ECESS. Further the modules can be easily adapted to new languages loading the language resources needed. This work has to be continued. Especially the not yet specified TC-STAR evaluation API has to be integrated.

Work on Text Processing Module

This work concerns the activities

- POS-tagging
- Grapheme-to-Phoneme Conversion
- Number handling

A tagger named synther [Sue03] has been developed, which was trained and tested on the WSJ Corpus using the Penn Tree bank POS system. With 1,2 Million word tokens for training and 20000 word tokens for test, a POS error rate of 3% for known words and of 11% for OOV words were achieved. This tagger has to be further optimized to decrease the error rate of OOV words and has to be extended to detect potential end of sentences.

For grapheme-to-phoneme conversion, the code based on neural nets was already developed in previous Papageno projects [Hai04]. Several tests have been performed for UK-English.

In order to handle numbers in the context of dates, weights, etc. language specific rules e.g. for UK-English will be developed. For this purpose a dedicated interpreter has been developed.

Work on Acoustic Synthesis Module

Work has started on investigating suited methods to select speech segments out of the voice corpus. It is planed to implement a triphone based concatenation search. Full work can start as soon the first voice corpus is ready.

4.3 UPC

The UPC Text-to-Speech system.

UPCTTS, the text-to-speech system from UPC is a multilingual system able to read text in several languages. The architecture of the system consists of a pipeline of modules that communicate through a rich data structure called multi-layer. This structure is able to represent the linguistic an acoustic properties of speech. Each module processes the input data and adds new information. The general principle is that modules code language independent technologies and the language dependencies are coded using external data. In most of the cases the information in the external data is derived automatically from data but some specific parts, as word normalization uses knowledge-based rules. The system includes a development suite including interpreted language, plugging facilities and scripting, so that different system configurations can be easily evaluated. Furthermore, several high-level interfaces have been developed, including SAPI.4 and SAPI.5. The system comprises modules for text processing (mark-up language, tokenization, word normalization, POS tagging, phonetic transcription), prosody generation (phrasing, duration assignment, f0 contour derivation, energy contour) and acoustic synthesis (unit selection, concatenation and manipulation). The most relevant proposals found in the literature have been implemented, especially in the field of prosody generation.

Work on Text Processing Module

At the beginning of the project the system derives the phonetic transcription using a knowledge-based approach. A dictionary containing canonical transcription was used combined with a set of

rules for OOV words. During this year some work has been done on a data-driven based grapheme-to-phoneme algorithm. Several algorithms have been evaluated including CART based, Support Vector Machine and Finite State Transducers. In preliminary results, Finite State Transducers give the best performance. Further work is required to derive syllabic boundaries and lexical stress**.**

Work on Acoustic Synthesis Module

Our system is based on unit selection. Segments of the database are selected to match the acoustic and phonologic features of the required units and to reduce the concatenation mismatch. The basic unit is diphone but the phonologic costs push the algorithm to select triphones or even words. Furthermore, the concatenation cost pushes the algorithm to select longer units. At the beginning of the project the concatenation cost only used F0 and energy parameters but not spectral information. During this year this information has been added to the concatenation cost.

The concatenation and manipulation module is based on PSOLA. In fact, in most of the cases, the PSOLA algorithm is only used to concatenate the units because usually the units found in the database have acoustic features which are similar to the ones needed. Till now the concatenation point between two segments was defined by hand, during the labelling of the speech database. In TC-STAR we have decided to segment part of the database automatically and include only the phoneme boundaries (not the concatenation point). Therefore, we have implemented an algorithm to select automatically the concatenation point. It consists on sweeping within some restrictions two adjacent units looking for the lower spectral distance between the two concatenation point candidates. This algorithm alleviated the problem on automatic segmentation. Perceptual experiments show that the quality is the same than the quality using hand-labelled concatenation points.

5. Voice conversion, manipulation and compression

This chapter describes the research progress on voice conversion, manipulation and compression.

Voice conversion aims at changing a reference voice into another given voice, allowing to customize a system to a given voice using few resources or to create a ‘unique’ corporate voice in many languages (cross-language voice conversion) [Sue05a]. The same technology is to be used at the segmental level to produce *expressive speech* [Kaw03] (cf. Chapter 6).

In particular, two issues are addressed: the speech representation and the transformation of parameters describing speech. This leads to two further activities within the scope of the TC-Star work package 3: fundamental research on *speech generation models* and on *speech manipulation* with very low degradation levels. E.g., speech features as prosody, frequency of voicing, formant position, or spectral tilt are to be manipulated without affecting the speech quality. Deeper insight into the generation of speech often leads to surprising approaches for *speech compression* which is another WP3 activity. Here, the main goal is to achieve high compression factors while preserving the speech quality.

Over the last twelve months, in particular, the following issues have been focused on by the involved partners (in doing so, they used their own preexisting language resources since the TC-Star speech corpora are not yet available):

- Nokia’s research on voice conversion has been based on parametric modeling of speech. During the first part of the project, the research has mostly focused on the conversion of the vocal tract contribution using a linear transformation derived from a GMM representation

of the source and target speech parameterized as LSFs. The conversion of the excitation part has also been tentatively studied. The results achieved so far have been very promising but additional research will still be needed to reach the full potential of the parametric approach. (Nokia)

- A recently published approach to text-independent VTLN-based voice conversion was extended to be applicable to GMM-based voice conversion training. This is important since several applications require the training data of source and target speaker to be non-parallel. Besides, in the future, this technique is to be used for cross-language voice conversion as well. However, up to now, text-independent approaches still significantly degrade the quality of the converted speech. (Siemens)
- When using GMM-based voice conversion, the spectral envelope is parameterized (to LPC, LSF, or MFCC) and compressed to only a few parameters. After conversion, the parameters are transformed back to time or frequency domain and concatenated by means of PSOLA techniques. It is obvious that due to the strong signal compression (e.g. from 200 samples of a certain time frame to 16 LSF coefficients), a lot of spectral details are lost resulting in a harsh signal quality. As former publications suggested, this problem can be overcome by predicting the residuals of the converted time frames. Several residual prediction techniques proposed in literature were compared with a novel approach based on a time-variant smoothing. It turns out that the smoothing technique outperforms the others in terms of conversion performance and speech quality. (Siemens)
- State-of-the-art voice conversion is based on a GMM describing the probability distributions of parameter vectors which represent the spectral information of the considered time frames. In operation phase, the coefficients of the GMM describing a time frame's properties are converted to the target coefficients by applying a linear transformation. This is done independently for each frame, i.e., without taking preceding or subsequent frames into account. As it seems that dynamic characteristics play an important role when a speaker's identity is to be converted, the GMM-based system is extended to an HMM-based system using dynamic features. (UPC)
- Up to now, most of the voice conversion systems use only spectral information coded as parameter vectors. For conversion, a linear transformation is applied to these vectors. In the novel conversion system, the effects of including phonetic information in the training of the mapping function and of the transformation were investigated. In particular, the phoneme type, a voiced/unvoiced flag, the point of articulation, manner and voicing were considered. The training was carried out in an unsupervised way using CART decision trees. (UPC)
- Currently, for high quality speech synthesis, a large amount of resources is required. For the use of speech synthesis in embedded systems, like mobile phones, toys, etc., this cannot be accepted. For such applications, the memory and time consumption has to be strongly reduced. In general, the largest part of the memory is required by the speech inventory. The inventory compression is carried out in three steps. At first, the number of stored speech segments (diphones, triphones, etc.) is limited, e.g. variants or infrequent segments are removed resulting in a baseline footprint of about 5MBytes for a diphone inventory @ 16Bit, 16kHz. Then, the quantization or the sampling frequency are reduced, e.g. to telephone quality 8Bit, 8kHz. Thirdly, the speech is encoded using for instance adaptive multi rate (narrowband / wideband). (Siemens)
- The current work on speech generation models is focused on extending speech models with appropriate parameters representing global speech features. This aims at manipulating the parameters in order to generate high quality speech signals with certain desired features. The main focus is to find an analyzing technique which extracts these speech features for manipulating and inversely synthesizing. Cepstral and sinusoidal models seem to feature a high potential in this area. (Siemens)

- In order to explore new models for speech production and manipulation, several tools were developed. In particular, more complex parameter representations resembling the physical production system were investigated e.g. taking the glottal source and the vocal tract characteristics into account. A tool for speech analysis/synthesis has been developed using the deterministic + stochastic model (sinusoids + noise). Different methods for stochastic component extraction, analysis and synthesis and for deterministic component synthesis have been tried and compared. (UPC)

In the following sections, the partners' contributions are described in more detail.

5.1 Nokia

Nokia's research on voice conversion has been based on parametric modeling of speech. During the first part of the project, the research has mostly focused on the conversion of the vocal tract contribution but the conversion of the excitation part has also been tentatively studied. The results achieved so far have been very promising but additional research will still be needed to reach the full potential of the parametric approach. Especially, the work on the excitation conversion must be continued as a part of future research.

Signal representation

To facilitate voice conversion, the speech signal is separated into vocal tract contribution and excitation signal. The separation is done using the well-known linear prediction (LP) scheme, i.e. using a source+filter model in which the source approximately corresponds to the excitation and the filter models the vocal tract. Furthermore, the excitation signal is represented using a parametric model. The parameters used in the model are pitch, voicing, gain (signal power) and the spectral representation for the excitation. A significant advantage of this model is that, in addition to its useful features from the viewpoint of voice conversion, it also lends itself to efficient compression.

Conversion approach

In our experiments, we have used the Gaussian mixture modeling (GMM) approach for the conversion. More specifically, the speech parameters are mapped using a locally linear transformation based on GMMs whose parameters are trained by joint density estimation. Combinations of aligned source and target parameter vectors, having the form $z = [x^T y^T]^T$, are used to estimate the GMM parameters (α, μ, Σ) for the joint density $p(x, y)$. The aim is to minimize the mean squared error $\varepsilon_{mse} = E\left[\|y - F(x)\|^2\right]$ between the converted source and the target speech. The conversion function F is

$$F(x) = E[y \mid x] = \int dy \cdot y \cdot p(y \mid x) = \sum_{i=1}^{Q} h_i(x) \cdot \left[\mu_i^y + \Sigma_i^{yx} \cdot \Sigma_i^{xx^{-1}} \cdot (x - \mu_i^x)\right]$$

where Q is the number of mixtures and h_i denotes the posterior probability that the i^{th} Gaussian component generated x (using the Bayes formula). Moreover,

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \text{ and } \mu_i = \begin{bmatrix} \mu_i^{x} \\ \mu_i^{y} \end{bmatrix}.$$

The parameters of the conversion function are unconstrained by allowing full covariance GMMs. It is also possible to perform simplified conversion with lower computational load by making the approximation that both Σ_i^{xx} and Σ_i^{yx} are diagonal. However, based on our experiments, it is usually beneficial to use full matrices.

Conversion of the vocal tract contribution

The vocal tract contribution is approximated using the linear prediction coefficients. For 8 kHz narrowband signals, we have used the LP order of 10 and the coefficients are estimated using the autocorrelation method. Before the actual conversion, the LP coefficients are converted into their linear spectral frequency (LSF) representation. This representation is used because of its favorable properties such as the guaranteed stability of the converted LP filter.

As discussed above, the conversion is done using the GMM approach. The source parameter vectors (LSFs) are converted using the conversion function *F* with parameters from the trained GMM. The selection of the number of mixtures has a direct effect on the conversion quality. According to our experiments, a training set of nearly 300 sentences can be effectively modeled by an 8-mixture GMM. The exact effect that the size of the training set has on the output is still under study.

The performance of this vocal tract conversion approach is dependent on successful training of the GMM model. More information on this aspect is given in the subsection called “Training of the conversion model”.

Conversion of the excitation signal

The most important excitation parameter, from speech perception point of view, is pitch. Consequently, we have first focused on this parameter. The parametric model allows very high quality pitch modifications and thus the main problem is to develop a good conversion technique. In our experiments, we have adjusted the pitch of the source speaker’s residual using a separate GMM and also using a single Gaussian to match the target’s pitch in average and variance. The results show that in practice a simple Gaussian model can perform the pitch conversion with similar quality as a 5-mixture GMM.

We have achieved very promising results by converting only the vocal tract contribution and the pitch but, in order to achieve even better quality, it will be necessary to also convert the other parameters (gain, voicing and the excitation spectrum). The research work related to these issues is ongoing. Especially, the conversion of the excitation spectrum requires additional research since the conversion cannot be done in a straightforward way due to the variable dimension of the spectral vector. Some tentative experiments have already been made but there are no concrete results yet.

Training of the conversion model

As training data, we have used speech material containing identical sentences from the source and the target speakers. Before training, the speech signals used as training material are aligned in time. The alignment procedure contains two steps. The first step is to generate phoneme-level labels for

the sentences. In the second step, the actual alignment is done using interpolation in the parametric domain. The alignment is based on the phoneme boundaries labeled during the first step and on the results of the dynamic time warping (DTW) algorithm.

At a more detailed level, the phoneme-level labeling is done automatically using Hidden Markov Model (HMM) based acoustic modeling. In the dynamic time warping step, we have made experiments by using different parameters in the alignment process: we have tried to use both line spectrum frequencies and Mel frequency cepstral coefficients (MFCCs). Further refinement of the alignment process is still ongoing.

Once the parametric training materials have been properly prepared, the GMM training can be handled using conventional training techniques. We have used a K-means type training approach and the expectation-maximization (EM) algorithm [Dem77]. In our experiments, the EM technique usually gave better results. In general, the training can be carried out successfully. However, in some rare cases numerical problems have been encountered in the training of the GMM based conversion model for the pitch parameter. This has happened with certain numbers of mixtures due to the fact that a non-positive definite matrix was fed to the Cholesky routine. Our current understanding is that the numerical problems can be avoided in the final implementation.

5.2 Siemens

5.2.1 Voice Conversion

Text-Independent Voice Conversion

So far, all conventional voice conversion approaches are text-dependent, i.e., they need equivalent training utterances of source and target speaker. Since several recently proposed applications call for renouncing this requirement, we developed an algorithm which finds corresponding time frames within text-independent training data.

The performance of this algorithm is tested by means of a voice conversion framework based on linear transformation of the spectral envelope. Experimental results are reported on a Spanish cross-gender corpus utilizing objective error measures, cf. [Sue04].

Automatic Segmentation

We are given the magnitude spectra of speech frames. These spectra are distributed among *K* well-distinct classes which can be regarded as artificial phonetic classes. This is done by clustering the spectra with the help of the k-means algorithm using the squared Euclidean distance as discrimination criterion. K-means delivers the class members as well as their centroid spectra $\overline{X}_k$.

Class Mapping

During training, we segment the given speech material of source and target speaker as described above. We get the source centroids $\overline{X}_k$ and the target centroids $\overline{Y}_l$. Now, for each target class l, we want to know the corresponding source class $k(l)$. When comparing spectral vectors of different speakers, it is helpful to compensate for the effect of speaker-dependent vocal tracts. This is done by using dynamic frequency warping and, afterwards, we are allowed to assess the similarity of two classes by means of the Euclidean distance:

$$k(l) = \arg \min_{\kappa=1,\ldots,K} D_{DFW}(\overline{X}_{\kappa}, \overline{Y}_{l})$$

.

Here, D_{DFW} is the distance between the frequency-aligned spectra derived from $\overline{X}_{\kappa}$ and $\overline{Y}_{l}$ by dynamic frequency warping.

Extracting Corresponding Time Frames

Once we have mapped one source cluster to each target cluster, we can shift the latter in such a way that each centroid $\overline{Y}$ coincides with the corresponding source centroid $\overline{X}$. Finally, for each shifted target cluster member $Y' = Y - \overline{Y} + \overline{X}$, we determine the nearest member of the mapped source class, *X*, using the Euclidean distance. The desired spectrum pairs consist of the respective unshifted target spectra *Y* and the determined corresponding source spectra *X*:

$$X = \arg \min_{\chi} \left| \chi - Y - \overline{X} + \overline{Y} \right| .$$

Parameter Training

Instead of using parallel training corpora aligned by dynamic time warping, we apply the above described mapping algorithm to non-parallel utterances of source and target speaker and proceed with the conventional linear transformation parameter training (expectation-maximization algorithm for Gaussian mixture models).

The Experimental Corpus

The corpus utilized in this work contains several hundred Spanish sentences uttered by a female and a male speaker. The speech signals were recorded in an acoustically isolated environment and sampled at a sample frequency of 16 kHz.

Objective Evaluation

As objective error criterion we use the relative spectral distortion D which compares the distance between the converted speech (represented by the vector sequence $\widetilde{x}_1^N$) and the reference (y_1^N) with that between source (x_1^N) and reference:

$$D = \frac{\sum_{n=1}^{N} d(\widetilde{x}_n, y_n)}{\sum_{n=1}^{N} d(x_n, y_n)} .$$

The following table shows the performance of both training methods based on parallel and on non-parallel training data in terms of the above defined objective error measure.

		D
male-to-female	text-dependent	0.39
	text-independent	0.47
female-to-male	text-dependent	0.38
	text-independent	0.49

Table 5.1 Objective comparison between text-dependent and text-independent voice conversion

Residual Prediction for Voice Conversion

Several well-studied voice conversion techniques use line spectral frequencies as features to represent the spectral envelopes of the processed speech frames. In order to return to the time domain, these features are converted to linear predictive coefficients that serve as coefficients of a filter applied to an unknown residual signal. In our work, we compare several residual prediction approaches that have already been proposed in the literature dealing with voice conversion. We also describe a novel technique that outperforms the others in terms of voice conversion performance and sound quality, cf. [Sue05b].

Residual Selection

The residual selection technique stores all residuals r_n seen in training into a table together with the corresponding feature vectors v_n that this time are composed of the line spectral frequencies and their deltas [Ye04].

In operation phase, we have the current feature vector $\tilde{v}$ of the above described structure and choose one residual from the table by minimizing the square error between $\tilde{v}$ and all feature vectors seen in training ($S(v)$ is the sum over the squared elements of a vector v):

$$\tilde{r} = r_{\tilde{n}} \text{ with } \tilde{n} = \arg\min_{n=1,\ldots,N} S(\tilde{v} - v_n) .$$

A Novel Approach: Residual Selection and Smoothing

The novel technique described in this section is an integral approach that tries to simultaneously handle inaccuracies of the residual selection and phase prediction as well as the treatment of unvoiced frames by means of a time-variant residual smoothing.

We are given the sequence $\tilde{r}_1^K$ of predicted residual target vectors derived from the formula in the previous paragraph, a sequence of scalars σ_1^K with $0 < \sigma_k \leq 1$ that are the voicing degrees of the frames to be converted, and the voicing gain α.

At last, we obtain the final residuals by applying a normal distribution function to compute a weighted average over all residual vectors $\tilde{r}_1^K$, the deviation is defined by the product of voicing degree and gain:

$$r_k^* = \sum_{\kappa=1}^{K} N(\kappa \mid k, \alpha\sigma_k) \cdot \tilde{r}_\kappa .$$

This equation can be interpreted as follows: In case of voiced frames ($\sigma \approx 1$), we obtain a wide bell curve that averages over several neighbored residuals, whereas for unvoiced frames ($\sigma \to 0$), the curve approaches a Dirac function, i.e., there is no local smoothing, the residuals and the corresponding phase spectra change chaotically over the time as expected in unvoiced regions [Sue05b]. This residual is normalized to have the same energy than before the smoothing.

Subjective Evaluation

The goal of the subjective evaluation of the described residual prediction techniques was to answer two questions:

- Does the technique change the speaker identity in the intended way?
- How does a listener assess the overall sound quality of the converted speech?

We want to find the answers by means of an extended ABX test and a mean opinion score (MOS) evaluation.

Now, 10 evaluation participants were asked if the converted voice sounds similar to the source or to the target voice or to neither of them (extended ABX test). Furthermore, they were asked to assess the overall sound quality of the converted speech on an MOS scale between 1 (bad) and 5 (excellent). Table 5.2 reports the results of the extended ABX test and Table 5.3 those of the MOS rating depending on the residual prediction technique and the gender combination. The methods not included in this section are described in [Sue05b].

%	source	target	neither
source residuals	20	10	70
reference residuals	0	79	21
codebook method	0	70	30
residual selection	0	70	30
selection & smoothing	0	**85**	15
selection* & smoothing	0	80	20

Table 5.2 Results of the extended ABX test

	m2f	f2m	total
source residuals	3.2	3.7	3.5
reference residuals	3.0	3.0	3.0
codebook method	1.6	1.9	1.8
residual selection	1.7	2.3	2.0
selection & smoothing	2.2	2.9	2.6
selection* & smoothing	2.2	2.8	2.5

Table 5.3 Results of the MOS test

Conclusion

We compared several residual prediction techniques to be used for voice conversion. The presented residual selection technique with smoothing outperforms the others in terms of voice conversion performance and speech quality. However, subjective tests show that, in general, voice conversion

still perceptibly deteriorates the quality of the source speech whereas most of the compared techniques succeed in converting the speaker identity.

5.2.2 Speech Generation Models

The current work is concentrated on further developing speech models with appropriate parameters representing main speech features. The goal is to manipulate the parameters in order to generate high quality speech signals with desired features. The main focus is to find an analyzing technique which extracts these speech features for manipulating and inversely synthesizing. Cepstral and the sinusoidal models seem to feature high potential.

The parameters of the cepstral model are filter coefficients which represents spectral characteristics of the vocal tract. The input of the filter is a signal with a flat spectrum. The output is a minimal phase speech signal. Manipulation of the base f0 is done by changing the pitch frequency of the input signal. Phoneme durations correspond to the rapidness of updating filter coefficients. Advances in speech signal quality seem to be possible by enhancement of the input signal.

The widely used Fourier transformation for analyzing speech segments causes several disadvantages. A concurrent analyzing technique is the sinusoidal modeling. It offers better parameters which are easier to manipulate. The extraction of the proper parameters is harder than the extraction by means of the Fourier transformation.

5.2.3 Acoustic Synthesis for Low-Footprint Systems

Currently, for high quality speech synthesis, a large amount of resources is required. For the use of speech synthesis in embedded systems, like mobile phones, toys, etc., this cannot be accepted. For such applications, the memory and time consumption has to be strongly reduced.

State-of-the-art systems for text-to-speech (TTS) conversion use speech segments from recorded natural speech. This method offers better naturalness compared to the parametric synthesis used in former systems. The segments are concatenated in time domain. As the concatenation points affect the quality of the synthesized speech, there is a tendency towards large segments by which the number of these points is minimized. Of course, this approach requires a large amount of speech data which occupies memory space of about 5 … 10 megabytes for a simple diphone inventory up to some gigabytes for larger corpora. It is necessary to compare these requirements to the resources available in embedded systems. Although they are growing with the technical progress in general, they are limited mainly by the expenses. There seems to be a “magic” border of 1 megabyte for the footprint of a TTS system as a whole, this means the program code as well as the databases including speech segments, rule systems, etc. [Sch02].

Inventory Compression

It is quite obvious that only the “smallest” of the concatenative TTS systems offers the chance to be shrinked to a footprint of 1 megabyte. This means that it will use a diphone inventory that limits the naturalness to a certain degree.

Starting point is the uncompressed diphone inventory with the size of approximately 5 megabytes at 16 kHz sampling rate and 16 bit quantization (256 kbit/s). Pitch markers and the descriptions of all units (diphones) are included in the inventory.

At a first step, the bandwidth could be reduced to telephone quality to adjust the application environment. With regard to halve the sampling rate the size reduction amounts to 50 % (128 kBit/s).

For further reduction, the following encoders are used and tested [Str03]:

1. **A**daptive **D**ifferential **P**ulse **C**ode **M**odulation (ADPCM)

   The ADPCM codec records the difference between adjacent samples of the signal and adjust the coding scale dynamically to accommodate large and small differences.

   Applying the codec to the speech units of the inventory, a reduction rate of 4:1 is achievable. The size of the resulting 8 kHz 4 bit-ADPCM inventory is about 800 kByte.

2. **A**daptive **M**ulti **R**ate-**N**arrow **B**and (GSM AMR-NB)

   The AMR-NB is a special speech codec and works on the principle of Algebraic Code Excited Linear Prediction (ACELP). It has eight basic bit rates: 12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15 and 4.75 kbit/s. At the encoder side, the speech signal frames are decomposed in an excitation signal and corresponding filter coefficients for spectral weighting. The decoder synthesizes the speech signal by filtering the reconstructed excitation with the coefficients calculated by the transferred line spectrum frequencies (LSF). According to the used bit rate, compression rates from 10:1 to 27:1 are achievable and result in inventory sizes of about 350 to about 170 kBytes.

3. **A**daptive **M**ulti **R**ate-**W**ide **B**and (GSM AMR-WB)

   The AMR-WB codec works in a similar manner as the AMR-NB with respect to the enlarged bandwidth of internal 12.8 kHz. The decoder output at 16 kHz is achieved by enhancing the signal at upper frequencies with noise estimated from the lower frequency bands. The following bit rates are adjustable: 6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05 and 23.85 kbit/s.

   According to the used bit rate, compression rates from 10:1 to 38:1 are achievable and result in inventory sizes of about 640 to about 210 kByte.

Acoustic Synthesis with Compressed Inventories

Acoustic synthesis consists of applying prosodic targets to selected speech units of the inventory and concatenating them. Due to the principle of some speech codecs, especially Linear Prediction (LP) based codecs like AMR, the acoustic synthesis is integrable in the decoder [Hof03]. The manipulation of the speakers' base f0 can be done on the excitation signal before the filtering step of the decoder. By deleting or doubling excitation frames, the target phoneme durations are manipulable at the excitation domain as well.

5.3 UPC

5.3.1 Voice Conversion

An already proposed, mapping function for vocal tract conversion is based on GMM as a model for joint source and target acoustic space. To estimate the GMM, aligned source-target feature vectors

are used. GMM-based systems work frame by frame, using only spectral information to learn the mapping and transform voices. Our study has been focused on the following points:

- The effects of including dynamic characteristics in the acoustic model used to build local vocal tract mapping functions. For this reason, GMM-based systems are extended to HMM-based systems, which can model dynamic characteristics.
- The effects of including phonetic information in the learning of the mapping function and in the transformation. The learning will be carried out in an unsupervised way by CART decision trees.

HMM-Based Vocal Tract Conversion

The block diagram of a HMM-based VC system is presented in Figure 5.1. In the training step, an HMM is estimated from LSF training vectors and then a conversion function is calculated for each state of the HMM. In the transforming step, the HMM is used twice. First, source data is segmented according the HMM states. Then, each frame is transformed applying the state-dependent conversion function.

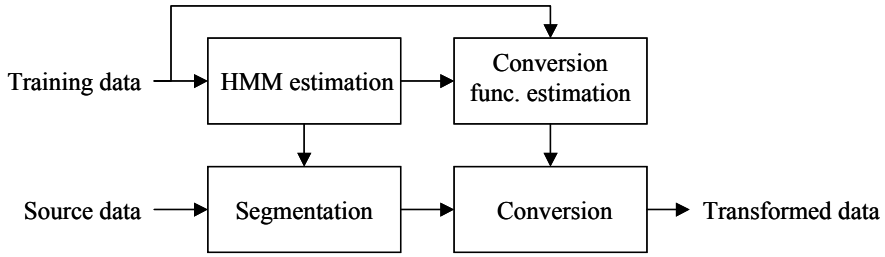

Figure 5.1 Block diagram of an HMM-based voice conversion system.

Source HMM-Based System

The basic idea of this system is to model the dynamics of the source speaker with an ergodic HMM. The steps for training the conversion function are the following. First, a source HMM is estimated from source data. Then, using the estimated HMM, source training vector sequences are segmented according to the optimal state path (using Viterbi search). All the vectors, with their target alignments, are collected for each state, and *N* (number of states) joint Gaussian functions are estimated. Finally, by regressing the function for each state, we have:

$$F(x) = \mu_s^y + \Sigma_s^{yx}(\Sigma_s^{xx})^{-1}(x - \mu_s^x)$$

This is a conversion function, where *s* indicates the state, **x** and **y** aligned source and target vectors, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ mean vectors and covariance matrices. To transform a new sequence, we need to segment it according to the source HMM. Then, the conversion function of each state is applied to each state parameter.

Joint HMM-Based System

As it has been previously done with GMM systems, we introduce joint information in order to allocate the distribution functions more judiciously, and also to use both source and target dynamic

information. So, using aligned source-target features vectors, a joint HMM is estimated. Like in joint GMM, there is no need of an extra step to calculate the mapping function for each state. Since there is a joint Gaussian per state, we can calculate their regression function straight forward. In the conversion phase, the best state sequence is the one that maximices the probability of the input and thee state dependent transformation $(x, F_s(x))$.

Decision Trees-Based Vocal Tract Conversion

Previous GMM-based systems work with spectral features to estimate the conversion function and to transform new source spectral vectors. The inclusion of phonetic information for each frame, such as the phone, a vowel/consonant flag, point of articulation, manner and voicing, was studied.

To estimate the mapping function, a CART decision tree has been used. The tree extracts, at each splitting step, overlapping regions of the acoustic space that can be represented by only one acoustic class, modeled by a joint probability function. The procedure to grow the tree is as follows. A GMM-based voice conversion system with one component is estimated from a training data set for the parent node (the root node in the first iteration), and an error index for all the elements of the validation data set is calculated. The error index used is:

$$E = \frac{1}{M}\sum_{m=0}^{M-1}\frac{D(y_m^{conv}, y_m)}{D(x_m, y_m)}$$

where x_m , y_m and y_m^{conv} are the source, target and converted *mth* frame respectively, and $D(\cdot)$ indicates an inverse harmonic mean distance.

Then, all possible questions of the form '*phonetic property* n=value' are evaluated and two child nodes are populated for each question. For each child node, a GMM with one component is estimated and the error index for the vectors of the validation set corresponding to this child node is calculated. The decision to let the tree grow is:

$$E_{parent} - \frac{(E_{child1} * elem_{child1}) + (E_{child2} * elem_{child2})}{(elem_{child1} + elem_{child2})}$$

where $elem_{child}$ indicates the number of spectral vectors of the validation set belonging to the child node. Only when this decision rule is positive and the number of training frames is higher than 25, this node is a candidate to be split with this question. At each iteration, the node with the decision rule with higher value for any question is split according to that question. The tree grows until there is no node candidate to be split.

To transform new source vectors, they are classified into leafs according to their phonetic features by the decision tree. Then, each vector is converted according to the GMM-based system belonging to its leaf.

Conclusions

We have evaluated the three explained vocal tract conversion systems and a GMM system by objective and perceptual criteria. According to objective criteria, when few training data is available, GMM, source HMM and CART systems perform in a similar way. But when the amount of training data increases, CART systems outperform GMM and source HMM. So, the inclusion of

phonetic information allows a better splitting of the acoustic space. Also, CART systems do not need any parameter tuning, a very computationally expensive part of GMM and HMM-based voice conversion. However, CART systems need training phonetically labeled training data, that restricts their applications.

Concerning the use of joint source-target information to estimate HMMs, from the experimental results it seems better to use only source data. We must take into account that using joint data increases the vector dimensions and there can be more inaccuracies from estimations with few training data.

The listeners reported that all the methods explained in this document succeed in changing the speaker identity. When they are asked about GMM and source HMM systems, they were not able to notice any difference. But, when GMM-CART pairs were compared, listeners preferred the CART system in 71% of the cases. These perceptual results are correlated with the objective results.

5.3.2 Voice Source Modeling and Voice Quality

We are currently developing tools to automatically perform source-filter deconvolution of the speech signal. We try to obtain useful parameters related to voice quality and a parameterized representation of both the glottal source and the vocal tract.

The main goal is to obtain a signal representation useful for performing expressive speech synthesis, voice conversion, and other applications requiring finer grain control of the voice properties. Traditional synthesis techniques based on concatenation of pre-recorded signals (e.g. TD-PSOLA, LP-PSOLA) are limited in the amount of modifications that can be performed. Thus, more complex representations resembling the physical production system are needed. We are using a simplified model of the human speech production system composed by two main blocks: the glottal excitation, consisting of a glottal source and additive aspiration noise, and the vocal tract filter, modeled here as an all-pole filter. This is somewhat unrealistic, since in reality there is physical interaction between the source and the vocal tract. With this simplification, the complexity of the synthesis model is reduced and several techniques can be used to estimate both the vocal filter and the glottal excitation.

Our current work in determining the parameters of this model is based on inverse filtering. If we know which filter was used, we can inverse-filter the speech and obtain the glottal excitation. As we do not know the filter, we need to make some assumptions. We use a widely accepted mathematical description of the glottal excitation (KLGLOTT model), and perform a joint estimation of its parameters together with the filter coefficients.

We are trying different approaches to obtain the parameters. So far, the technique giving the best results is to approach the estimation problem as a convex optimization procedure, where the mean square error between the resulting synthetic speech and the original speech is reduced.

In order to improve the quality of the overall system, the glottal excitation is re-parameterized using a more realistic model (LF model). We are experimenting with different approaches to do this automatically. A direct way to determine the LF model's 4 parameters (3 time instants and the amplitude) is to perform an initial estimation by inspection of the inverse-filtered speech (noisy glottal excitation). Because of the noise, it is necessary to apply an optimization process to this initial estimation to reduce errors. We are working now with different techniques (e.g. gradient-descent methods, Newton algorithms, Simplex search) and different objectives (e.g. minimization of the root mean-squared error).

We are in the preliminary stage of using some of the techniques used in CELP speech coding to improve the quality of the synthetic speech (incorporating perceptual information to the minimization-error criterion, so that we focus on the most perceptually relevant features).

A tool to extract different standard voice quality measures from this model is under development. By means of these measures, we expect to be able to add useful information to the signal generation module of the speech synthesizer, in particular to synthesize expressive speech. Another application is to improve the quality of our current voice conversion system by adding the voice quality measures or the glottal parameters to the set of parameters currently used to perform the conversion.

5.3.3 Speech Analysis, Synthesis, and Manipulation Based on the Deterministic + Stochastic Model

A tool for speech analysis/synthesis has been developed using the deterministic + stochastic model (sinusoids + noise). Different methods for stochastic component extraction, analysis and synthesis and for deterministic component synthesis have been tried and compared. No significant differences have been found between them. The synthetic signals obtained are almost indistinguishable from the original when the phase information is kept. Synthesis without phase information has been studied, and a high quality of sound was reached. The block diagram of the system is shown in Figure 5.2.

Time-scale and pitch-scale transformations have been implemented, and several alternatives for phase reconstruction have been evaluated, by means of simple frequency linear interpolation or vocal tract estimation. The quality of the transformed synthetic signals is high, but it is still worse than the quality achieved by PSOLA methods.

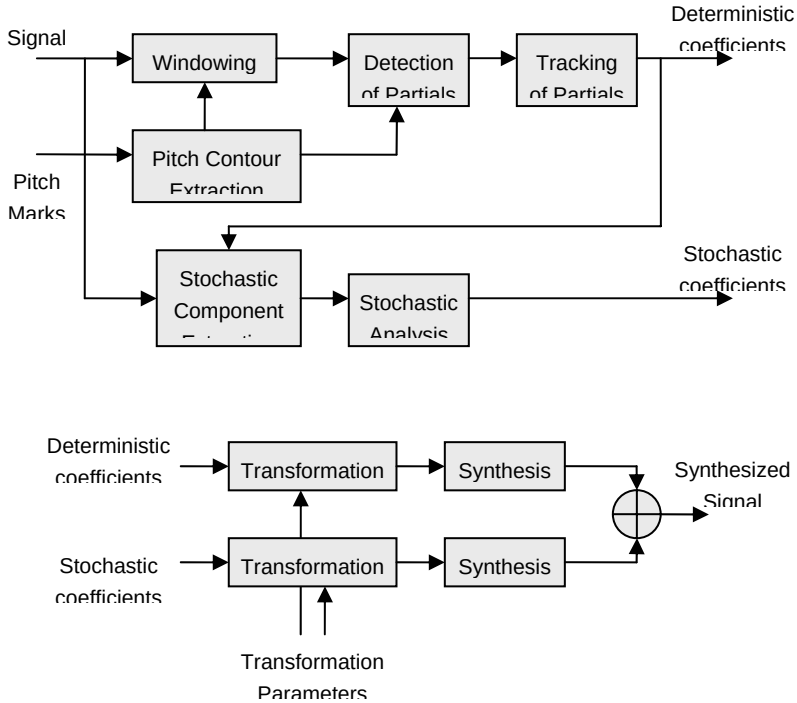

Figure 5.2 Block diagram of the analysis/modification/synthesis tool.

Concatenation of parameterized speech fragments following the deterministic + stochastic model is currently being investigated. Future work will be focused on voice conversion between different speakers and reducing the amount of training data necessary to perform the conversion.

6. Prosody modeling and expressive speech

6.1 Introduction and general framework

The work reported in this section has been done by UPC. Other partners have done some work on speech prosody but it has been reported in section 4, baseline systems for research and evaluation. In particular, a new approach has been proposed by Nokia to model prosody of Mandarin and other tonal languages using what has been named as eigen pitch.

In this section we report the general framework and the progress done to improve the expressivity of the speech synthesis in a speech-to-speech translation system. Usually speech synthesis focus on the linguistic information. However, using expressive speech much more information can be provided. For instance, the same words can have different, event contradictory meanings, depending on the intonation. Speech is able to give a lot of information from the speaker, for instance which is his emotional state or his/her attitude towards the message or towards the audience. However, convert text into expressive speech is a very difficult task. First of all, because some of the information, as the emotional state, is not included in the text. Sometimes the information is in the text (for instance the focus in a new message, or the typical speaker attitude towards certain information) but it cannot be deduced with today in unrestricted domains using today semantic and pragmatic models. Expressive speech is important in many applications of speech synthesis, for instance entertainment and education. In particular, it is very important in speech-to-speech translation because the source of the information is a speaker, not a writer. Therefore, speech contains this paralinguistic information that should be deliver to the listener.

In TC-STAR the key idea is to use information derived from the source speaker. For instance, if the speakers reveals surprise, this information should be preserved in the target speech. Of course this is a very ambitious and complex goal: how surprise is produced in speech may be language and event speaker dependent. Furthermore, defining the number of states (as surprise) and the degree is a very complex task. Instead we try to infer information about acoustic prosody and map the input prosody into the target speech. Several features are derived from the input speech using prosodic extraction algorithms. These features are mapped into the target text using the alignment information provided by the automatic translation system. In this way, one or several words in the source text may be mapped onto one or several words in the translated text. This new features are added to the linguistic information (i.e. the translated text) to produce synthetic speech. Figure 6.1 shows the architecture of a speech to speech translation system. Figure 6.2 shows the architecture of a speech to speech translation system with embedded *prosody translation* [Ekl95].

The *prosody extraction* module inserts additional information into the output text of the speech recognition module. Then, the *transfer* (*prosody mapping*) performs the transformation of the input prosodic annotation into the output text of the text translation module taking into account alignment information. Finally, the speech synthesis module produces the output waveform signal using the prosody generated by the *prosody generation* module, which takes advantage of the enriched input text.

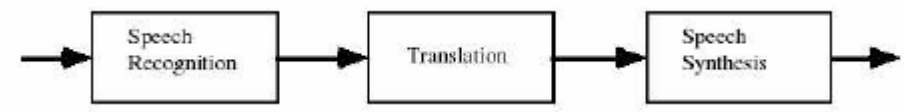

Figure 6.1: Architecture of a speech to speech translation system

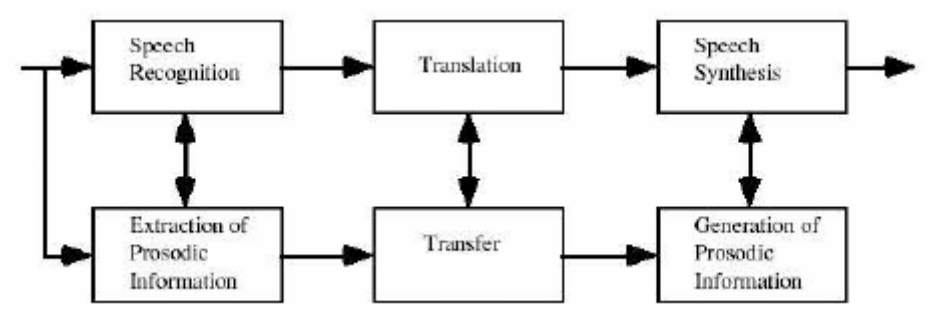

Figure 6.2: Architecture of a speech to speech translation system with embedded prosody translation

Some preliminary experiments have been done to use the extracted information in speech synthesis, using a bilingual corpus, Spanish-Catalan from the tourist domain. The results are promising as information from the source is one of the main features used to derive the intonation.

During the first year the main work has been on generation of prosodic information: in order to include new features it is required that the prosody models are data driven and robust. Next section reports on the studied methods for phrase break prediction and intonation prediction which are two of the main prosodic features. Then, section 6.3 report some work on extraction of prosodic information.

6.2 Generation of the output prosody

This section reports on the studied methods to predict phrase breaks and intonation from text. In the next phase of the project, this methods will be extended to include other features derived from the source. To improve the performance of intonation models, a new training paradigm has been introduced. The idea is to use information from all the corpus to analyze each sentence. This can be seen as a top-down approach to intonation and provides robust estimations and better prediction results. This paradigm has been applied to two existing methods (Fujisaki and Tilt) and to a new one based on Bézier curves producing better objective and subjective results.

6.2.1 Phrase break prediction

In the literature several approaches have been proposed that use machine learning techniques to predict phrase breaks. In general, the methods consist of taking a decision after each word about whether a phrase break boundary must be placed or not. The decision considers context information and in some cases the previous decisions about the presence or absence of phrase break boundaries. Phrase break prediction is a difficult task because it highly depends on the meaning of the sentence, speech rate, domain, etc. In our work we explore several approaches to predict phrase breaks. We have proposed new methods or extended some published ones to compared the performance of the methods and to be able to include afterwards the new feautures [Bon04].

CART. This baseline method consists of the prediction of phrase breaks using classification trees and was proposed by Prieto et al. [Pri96]. The input features are: a 4-word POS window (POS: part of speech, morphological category of the word); 2-word window for pitch accents; the total number of words and syllables in the utterance; the distance of the word from beginning and end of the sentence in words, syllables, and stressed syllables; distance from the last punctuation in words;

whether the word is at the end, within, or at the beginning of an NP (Noun phrase), and if within an NP, its size and the distance of the word from the start of the NP..

CART-LM. One weakness of previous method is that decisions are made locally, without taking into account previous or next decision. But it is evident that the phrase breaks influence in the speech rhythm and phrase breaks are not independent. In fact, Prieto et al. [Pri96] included previous decisions in the next decision. In this extension we include the CART estimation not to do a hard decision but as the probability of having a phrase break. This information is combined with the probability of a break knowing the decisions made in previous words. This is modeled using language-model (LM) techniques. The Vitebi algorithm is used to find the best phrase-break positions given this two information souces.

CART-LB. In this method, the duration in words of each major phrase is modeled explicitely, not using a language model (n-gram). The search is performed using a level building algorithm that iteratively find the best solution of including *k* phrase breaks. This method not only allows to obtain the best position of the phrase break boundaries, but also to select the frequency of phrase breaks. This can be selected according to the desired speech rate.

Table 6.1 summarizes the results of the different algorithms using the F measure (which combines precision and recall).

	F global	F $\neg P$	F P
CART	80.13	80.04	79.92
CART-LM	82.65	70.43	70.45
CART-LB	87.32	81.05	81.02

Table 6.1: Summary of F-measure for each method

6.2.2 Intonation model

JEMA: Joint Extraction and Modeling approach.

The intonation model is an important component of text-to-speech systems which generates a suitable fundamental frequency contour for a given synthetic utterance. The intonation model training consists of two-stages: parameterization and rule inference using machine learning techniques.

The parameterization of the fundamental frequency contour permits a better generalization for the machine learning techniques of the second stage. For example, the parameterization enables to extrapolate to cases where the duration of the sentence is different. Then, the rule inference using machine learning techniques finds a mapping between linguistic features extracted from the utterance and the parameterization of the fundamental frequency contour. Once the intonation model is trained, the application of the inferred rules to linguistic features of new sentences enables to obtain a set of parameters to synthesize a suitable fundamental frequency contour.

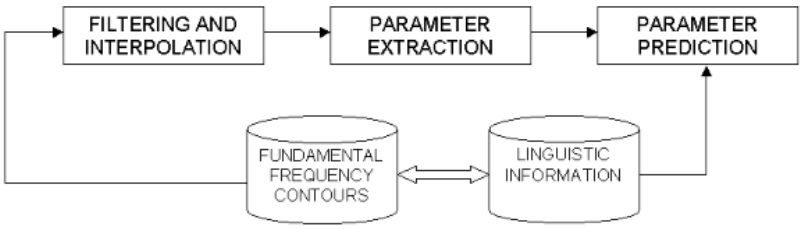

Figure 6.3: Two-stage approach

Most of the data-driven intonation models are estimated using two stages. This procedure presents some characteristics that can cause some training problems:

- **Interpolation of fundamental frequency contour.** An initial interpolation of f0 in the unvoiced regions is required. As this interpolation is somehow arbitrary, this may introduce *noise* in the extracted parameters: contours with the same F0 contour in the voiced parts may be represented by different parameters. This introduces dispersion in parameters reducing the accuracy of the machine learning techniques.
- **Multiple solutions.** In some intonation models, different values of the parameters can represent the *f0* contour with the same accuracy (e.g. mean square error, *MSE*). Again, this increases the variance of the parameters and reduces the accuracy of the machine learning techniques.
- **Sentence by sentence extraction.** Sentence by sentence parameter extraction lacks general information about the intonation of the language. The f0 contour is noisy, in the sense that it is affected by micro-melody, errors in the measure, etc. To solve this the f0 contours are usually filtered before computing the parameters of the model. But this filtering is somehow arbitrary. The knowledge of all the other sentences could be used as *a priori* probability to derive the underlying parameters.

The *JEMA* combines parameter extraction and model generation into a single loop [Agu04c]. The model generation is performed using machine learning techniques that cluster segments of F0 contours from the training databases. Each cluster is considered as a class that is approximated by a set of parameters given the intonation model, e.g.: Tilt parameters. The parameter extraction is performed using an optimization algorithm that finds the global parameters that best represent all the contours of the cluster. As the same parameters are used for all the contours, stylization or unvoiced interpolation is not required. The extracted parameters are more consistent and the prediction capabilities of machine learning techniques that are used to generate a model are improved. Figure 6.4 shows the scheme of joint approach. Further details of the application of *JEMA* to different intonation models are given in the following sections.

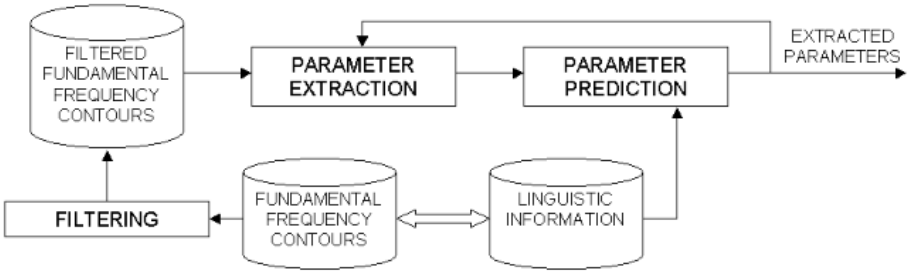

Figure 6.4: JEMA approach

Bézier intonation model.

Escudero et al [Esc02] proposed a phonetic representation based on Bézier curves using the accent groups in Spanish as the phonological unit. Bézier curves are based on a polynomial function where its coefficients allow a representation that is more meaningful than the resulting polynomial coefficients in expanded form. This parameterization was proposed by Escudero in his PhD. thesis due to several reasons:

- **Representation capacity.** The values of the coefficients are representatives of different portions of the contour.
- **Homogeneity of the representation.** Curves with different duration but same shape have the same set of parameters. In this way, this representation has properties of elasticity.
- **Tunable accuracy.** Increasing the order of the approximation reduces the approximation error.
- **Restrictions.** It is possible to restrict the shape of the approximation, e.g: continuity of order 0 and 1 to obtain smooth resulting contours.

The polynomial formulation is shown in equation 1 and the shapes of the base polynomials for a fourth order curve are shown in Figure 6.5. Bézier coefficients allow a meaningful representation compared with the final polynomial coefficients, which are more sensitive.

$$P(t) = \sum_{n}^{N} \alpha_n \begin{pmatrix} N \\ n \end{pmatrix} t^n (1-t)^{(N-n)} \tag{6.1}$$

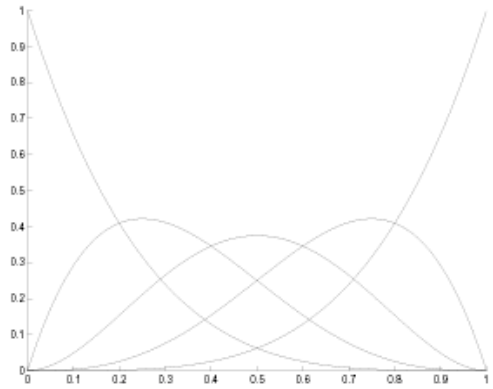

Figure 6.5: Bézier polynomials

Figure 6.6 shows an approximation of a fundamental frequency contour using Bézier curves for accent groups, with continuity constraints up to the first derivative.

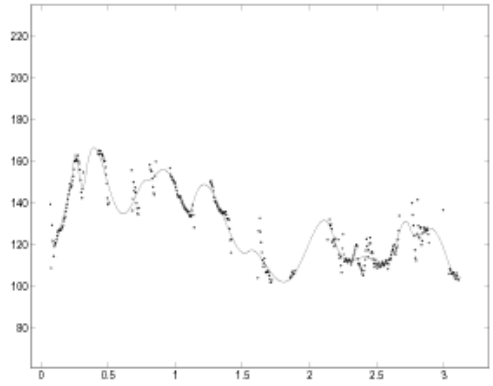

Figure 6.6: F0 contour approximated usingpiece-wise Bézier curves with five coefficients

JEMA approach.

The joint optimization framework imposes that the formulation to extract the optimal polynomial coefficients is modified [Agu04b]. The optimization is performed minimizing the mean squared error, but taking into account that:

- The error that is minimized is the global mean squared error.
- Two components are combined using Bézier curves: the intonation contour is made up from additive components, one for each major phrase and one for each accent group.
- The group of coefficients corresponding to a Bézier curve depend on a vector which maps minor phrase or accent group classes with positive integers (class number).

The mathematical formulation is shown in equation 6.2.

$$F_0^k(t) = \sum_i^{N_{MP}^k} P_{MP}^{C_{MP_i}^k}(t_{MP_i}^k(t)) + \sum_j^{N_{AG}^k} P_{AG}^{C_{AG_j}^k}(t_{AG_j}^k(t)) \tag{6.2}$$

where:

N_{MP}^k is the number of minor phrases of the k_{th} sentence.

N_{AG}^k is the number of accent groups of the k_{th} sentence.

$t_{MP_i}^k(t)$ is the temporal axis of the i_{th} minor phrase of the k_{th} sentence.

$t_{AG_{ji}}^k(t)$ is the temporal axis of the j_{th} accent group of the k_{th} sentence.

$C_{MP_i}^k$ is the number of the minor phrase class assigned to the i_{th} minor phrase of the k_{th} sentence.

$C_{AG_{ji}}^k$ is the number of the accent group class assigned to the j_{th} accent group of the k_{th} sentence.

In this function, P_{MP} and P_{AG} are the Bézier curves of the minor phrase and accent group components, respectively. Each curve has its own associated time axis, $t_{MP}(t)$ and $t_{AG}(t)$. The time axis range is zero to one. These curves are zero elsewhere.

The joint cost function is shown in equation 6.3. The goal is to minimize the mean squared error. This equation has a unique analytical minimum that is found using a set of linear equations.

$$e = \sum_{k}^{N_s} \left(\sum_{t}^{T_k} \left(f0^k(t) - \left(\sum_{i}^{N_{MP}^k} P_{MP}^{C_{MP_i}^k}(t_{MP_i}^k(t)) + \sum_{j}^{N_{AG}^k} P_{AG}^{C_{AG_j}^k}(t_{AG_j}^k(t)) \right) \right)^2 \right) \quad (6.3)$$

where:

N_s is the number of sentences.

T_k is the duration of the sentence.

Model training process

The idea behind the training process is to find a set of minor phrase and accent group clusters (obtained using linguistic information) that are optimal in the sense of mean squared error and Pearson correlation coefficient. Mean squared error and Pearson correlation coefficient are chosen as the optimization indexes because there is a common consensus on intonation modelling about using them to measure the prediction accuracy.

There are many ways to perform a clustering based on a set of parameters. Classification and regression trees [Bre84] are chosen, because of the capabilities to classify using continuous and discrete features. The information provided by the final tree can be valuable for future improvements or to get an insight of the main features related to the problem.Because of the superpositional approach, two independent trees are trained (accent group component tree and minor phrase component tree), with a joint optimization cost (Pearson correlation coefficient).Initially, each tree has a unique root node. As a consequence, there is only one minor phrase and accent group class. The steps performed to grow the trees are:

- Consider each possible splitting for each tree, according to linguistic parameters extracted from text.
- Find the optimal polynomial coefficents (α's and β's associated to minor phrases and accent groups) for each splitting.
- Select the split which maximizes the Pearson correlation coefficient.

The trees are grown until the Pearson correlation coefficient gain is less than a predefined threshold. The number of elements in each leaf is bounded to be superior than a predefined threshold (in our experiments, this threshold is 40), in order to prevent a weak modeling of cluster due to small data size.

The linguistic features used to predict minor phrases are: sentence type (declarative, interrogative or exclamative), number of minor phrase in the sentence, position of the minor phrase in the sentence, number of accent groups in the minor phrase, number of words in the minor phrase and number of syllables in the minor phrase.

The linguistic features used to predict accent groups are: sentence type (declarative, interrogative or exclamative), number of minor phrase in the sentence, position of the minor phrase in the sentence, number of accent groups in the minor phrase, number of words in the minor phrase, number of syllables in the minor phrase, number of following accent groups, number of accent groups in the sentence, number of syllables in the accent group and position of the accent group in the minor phrase.

The JEMA has a drawback that does not allow the definition of continuity constraints, because of the global nature of the problem. In the case of English, accent groups are defined as a sub-sequence of the sequence of syllables contained in a minor phrase, such that the first syllable is accented and the remaining syllables - if any - not accented [Spr98]. As a consequence, discontinuities in the fundamental frequency contour can be produced in an accent group boundary inside a word. This problem is overcome using a smoothing function in the boundaries of accent groups. This smoothing function performs a linear interpolation in the middle of the discontinuity.

Fujisaki intonation model

The Fujisaki's intonation model (Fujisaki et al. [Fuj84]) is based on a physical model of the fundamental frequency production system [Fuj00]. It is represented by two second-order filters. One filter is excited with pulses, and the other with deltas. The latter are related to phrase commands and pulses are related to accent commands. A DC value (*Fb*) is added to the output of these filters. A scheme is shown in Figure 6.7.

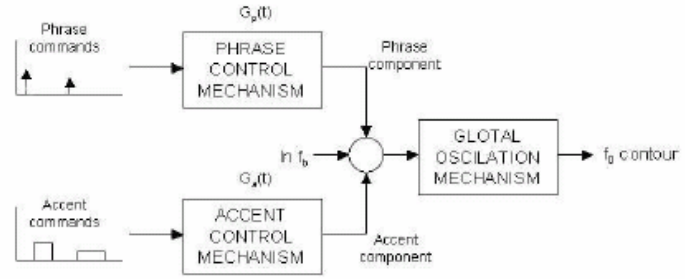

Figure 6.7: Fujisaki's model scheme.

Two-stage method

In our first experiments we used the classical two-stage approach for Fujisaki intonation modeling training. In this way, we obtain a base-line to compare JEMA approach.

In the first stage, command parameters are extracted sentence by sentence yielding optimal parameters for each contour of the training set. Next, the resulting parameters are used to train classification trees based on vector clustering. The main characteristic of the extraction procedure is the application of strong linguistic constraints:

- Each minor phrase is modeled by one phrase command. The phrase command can only appear within a window of 200ms centered at the beginning of the minor phrase.
- The number of accent commands inside an accent group is limited to one.

Each command is represented as one vector of parameters: $[A_P,T_0]$ for phrase commands and $[A_a,T_1,T_2]$ for accent commands. In the command prediction stage, a clustering of command parameter vectors is performed using regression trees. One of the trees is related to accent groups

(accent commands) and the other to minor phrases (phrase commands). The questions of these trees are related to the linguistic features of accent groups and minor phrases. The centroids of the clusters are the parameter vectors that minimize the mean distance to the other vectors of the cluster. In this way, possible deformations due to individual prediction of each command parameter are avoided.

Some observations led to improvements of the extraction procedure presented in [Agu04]:

- The default value α=3.0 for the phrase control mechanism was not optimal. In several experiments the global optimum value was found to be α =1.8 for the whole corpus.
- The window for accent command timing extends 50 ms the accent group boundaries. This extension overcomes the problems of F0 peaks that extend their influence outside the accent group.

Concerning command prediction, we found it is advantageous to predict accent command onset and offset, rather than to predict onset and duration. Compared to the results presented in [Agu04], these modifications resulted in better fitting accuracy as well as in better prediction performance. The synthesis capabilities of this method will serve as baseline for the evaluation of the new combined extraction and prediction algorithm.

Joint parameter extraction and prediction algorithm

In this section we propose a novel algorithm that applies JEMA approach to Fujisaki's intonation model training [Agu04d]. As in the previous method, two regression trees are grown using linguistic features as questions in the nodes. One tree is related to minor phrases, and the other to accent groups. As before, we assume that each minor phrase is modeled by one phrase command, and each accent group is modeled by one accent command. Thus, each leaf of the tree collects a set of fundamental frequency contours from the training corpus that must be approximated with command responses. A hill-climbing procedure is used to find the parameters that provide a global optimal approximation to all the fundamental frequency contours.

Due to the superpositional nature of the intonation model, each partition of one tree affects the optimal solutions of the parameters of the other tree. Therefore, the optimization must be jointly performed for phrase and accent commands. The steps of the algorithm are:

- Each tree (accent group tree and minor phrase tree) has an initial root node, which groups all the contours. An initial optimal solution is found that approximates all contours with the same phrase command for each minor phrase, and with the same accent command for each accent group.
- All possible questions are examined in the leafs. For each question, the optimal parameters for phrase and accent commands are determined, and the approximation error is obtained. The optimization is performed using a hill-climbing algorithm.
- The splitting questions for phrase and accent command trees are chosen. The selection criterion of the optimal node question is the minimization of the approximation error.
- Then, the global optimal values for α, β and F_b are searched using a grid of values.
- The process is iterated from the second step, until a minimum number of elements in the leafs is reached or the differential gain on accuracy is lower than a threshold.

The global optimization avoids the interpolation step of the stylization process, which can cause a bias in the parameter extraction. Another advantage of global optimization is the consistency of the

parameters. Non-consistent parameters increase the dispersion, and limit the prediction capabilities of machine learning techniques.

Tilt intonation model

Review of the Tilt intonation model

The Tilt intonation model defines intonational events which are represented by a set of curves. Intonational events are phrase breaks, accents, etc. The curves are piecewise and they are composed of a rise and a fall component. Each curve is connected with the adjacent curves by a line. The way the model approximates the fundamental frequency contour gives the name to Tilt parameters: rise, fall and connection (*RFC* parameters). In particular, the *RFC* parameters for a Tilt event are (see Figure 6.8): rise amplitude (A_r), rise duration (t_t), fall amplitude (A_f), fall duration (t_f), position (t_e) and F0 height ($f0_{offset}$).

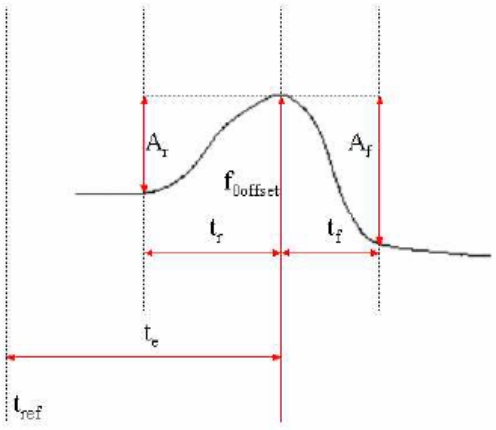

Figure 6.8. RFC parameters.

The intonation model requires detecting the Tilt events in the database, e.g.: phrase breaks events and accent events. This task can be performed using knowledge based rules or automatically derived rules. For instance, in [Dus00], HMM are used to detect Tilt events. After this preliminary process, the Tilt acoustic parameters are extracted from the sentences of the database (step 1). This task can be performed using gradient descent techniques or the method proposed in the next section.

In the second step, the linguistic information present in the sentences of the database is transformed into linguistic features and used for construction of the model. In this work, binary regression trees (CART) are used to predict the parameters of Tilt intonation events. The Tilt intonation model provides the mapping between linguistic features of the sentences and the parameterization of the fundamental frequency contour. Therefore it is possible to get linguistic features of new sentences and predict a suitable fundamental frequency contour.

The two-stage approach serves as a baseline for the evaluation of the novel algorithm based on *JEMA*.

Closed-form determination of amplitude parameters.

In Tilt intonation model it is not possible to obtain a closed-form solution for all the parameters of the model. However, it is possible to obtain the optimal solution for the amplitude parameters and $f0_{offset}$ assuming that the time instants are known.

The optimal values of the time instants can be found using grid search or gradient descent techniques. The time instants remain constant during closed-form amplitude optimization, and the amplitude values are kept constant during time instant optimization using gradient descent techniques. The update loop is shown in Figure 6.9.

We must point out that the loop that combines closed-form determination of some parameters and gradient descent of the other parameters has a better convergence rate. This optimization procedure is used both in the two-stage intonation model and in the joint approach presented in next section.

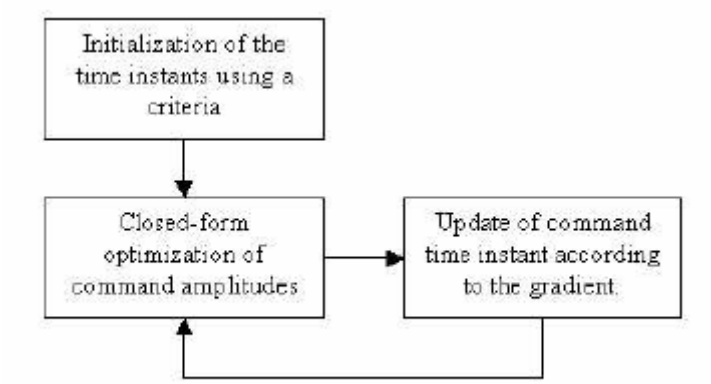

Figure 6.9. Update loop.

The closed-form formulation is obtained by minimizing the mean square error:

$$e^2 = (f0 - \hat{f0})^T (f0 - \hat{f0})$$

where $\hat{f0}$ is a function of all RFC parameters of each event:

$$\hat{f0} = f(A_r, t_r, A_f, t_f, t_e, f0_{offset})$$

In order to obtain the set of linear equations we take derivatives of the error

(t_r^i, t_f^i and t_e^i are kept constant):

$$\frac{\partial e^2}{\partial A_r^i} = 0 \quad \frac{\partial e^2}{\partial A_f^i} = 0 \quad \frac{\partial e^2}{\partial f0_{offset}^i} = 0$$

If this formulation is applied to the two-stage method, there is one set ($A_r^i, A_f^i, f0_{offset}^i$) for each event in the sentence. In the next section

the system of linear equations simultaneously finds the optimal solution for all the RFC parameters (A_r^i , A_f^i , $f0_{offset}^i$) of the training

corpus. Then the event i is used for all the contours which belong to the cluster i, defined by the machine learning technique (see the following section).

The gradient descent algorithm consists of an update equation (6) and (7). t_r^i, t_f^i and t_e^i are updated while A_r^i, A_f^i and $f0_{offset}^i$ are kept constant.

$$p_{n+1} = p_n - diag(\mu_n)\nabla e(p_n)$$

$$\nabla e(p_n) = [\frac{\partial e}{\partial t_r}, \frac{\partial e}{\partial t_f}, \frac{\partial e}{\partial t_e}]$$

Joint parameter extraction and prediction algorithm

In this section we propose a novel algorithm to apply the *JEMA* to the Tilt intonation model. The goal is to estimate simultaneously the Tilt parameters and the prediction model. As stated above, the global optimization avoids the interpolation step of the stylization process and produces more consistent parameters, improving their predictability from linguistic features.

Classification and regression trees (CART) are selected to estimate the model. The advantage is that they can use both discrete and continuous features. Furthermore, the representation provides useful information to increase the knowledge about the task. This information can be used for future improvements of the system. The classification tree is used to cluster the Tilt intonation events using questions concerning the prosodic and phonetic context of the events. Each leaf of the tree collects a set of fundamental frequency contours from the training corpus that must be approximated by the Tilt intonation parameters. The optimal parameterization is obtained using a combination of closed-form solution for the amplitudes and f0 offset and a hill-climbing procedure for time instants, as explained above. These optimizations provide a global optimal approximation to all the fundamental frequency contours in the training database.

The steps of the algorithm are:

1. The tree has an initial root node, which has to represent all the events. The initial optimal solution is found that approximates all contours with the same set of Tilt acoustic parameters.
2. All possible questions are examined in the leaves. For each question, the optimal Tilt acoustic parameters are determined and the approximation error is calculated. The optimization is performed using a combination of closed-form solution and hill-climbing algorithm.
3. The splitting linguistic question for the tree is chosen next. The criterion for selection of the best linguistic question splitting the node is the minimization of the approximation error.
4. The process is iterated (from 2) until a minimum number of elements in the leaves is reached or the differential gain on accuracy is lower than a threshold.

Results

Experiments where performed using a a female voice of a Spanish corpus of *500* sentences. The utterances were manually segmented in demiphones, and the fundamental frequency contour was obtained from the laryngograph channel. The train set for the experiments was *70%* of the corpus and the test set was *30%* of the corpus. The results are shown in Table 2.

The intonation models trained with the JEMA approach outperform two-stages approaches and intonation models based on hand-written rules.

Method	**RMSE [Hz]**	**Correlation**
Piece-wise linear	20.46	0.58
Bézier with JEMA	18.08	0.75
Two-stages approach	21.79	0.68
Fujisaki with JEMA	18.67	0.73

Table 6.2: Global results.

Other experiments were performed using an extended database with prompts recorded for a dialog system. The results are shown in Table 6.3. In this table we see that the three intonation models trained with the JEMA approach have similar performance.

Method	**RMSE [Hz]**	**Correlation**
Bézier	20.9	0.76
Fujisaki	21.2	0.76
Tilt	23.1	0.68

Table 6.3: Global results using extended database.

Objective measures are the first indicators about the performance of an intonation model. However, in order to have a measure of acceptance by final users a listening test has been performed by twelve subjects. They were asked to judge the naturalness of the intonation of several sentences using a five point scale (1:unnatural, 5:natural). Each intonation model predicts the F0 contour of the test sentences. This contour is imposed to the test sentences by resynthesis using Praat.

Table 6.4 shows the results of perceptual evaluation of naturalness for all methods. The natural intonation is included in the test as a reference for the evaluators and also to ensure the competence of the evaluators.

Method	**RMSE [Hz]**
Natural	4.6
Bézier	3.4
Fujisaki	3.5
Tilt	3.4

Table 6.4: MOS for 3 different intonation models trained with JEMA.

The table shows that the three intonation models are performing similar with MOS scores around 3.5. It also shows that we are far from obtaining the quality of natural contours. We believe that these results show that the main limitation is due to the prediction. More effort should be devoted to derive new features from the input text (including syntactic and semantic features) that may influence the intonation.

6.3 Analysis of the input prosody

The analysis of the prosody of the input voice is performed extracting numerical and symbolic representations that enable to perform the transfer of such information onto the target text and onto the output speech synthesis. The mapping task takes into account alignment information provided by the statistical spoken translation system.

6.3.1 Symbolic representation of the input fundamental frequency contour

The mapping of the pitch contour will be performed using a symbolic representation of the source contour. This symbolic representation is an abstract representation of the input contour that will be used to generate the output contour.

The input fundamental frequency contour is converted to a symbolic representation using clustering techniques that allow creating classes of contours. In order to produce expressive speech we want to find information which is not related to linguistic information: this information is available through the target text. Therefore, the cluster has to be done using acoustic information. The elements to populate the clusters are F0-contours of the accent groups. After the clustering of the source contours and using the alignment source text – target text, each accent group of the output language is linked to zero, one or more source accent-groups classes. This information is used as input features to train an intonation model. In this way, multiple information sources are being used to generate the output fundamental frequency contour. We can not rely only on the linguistic information of the translated text or on the information obtained from the input voice. Each information are complementary and in some cases even contradictory. In this way, the machine learning technique used to infer an intonation model finds out the regularities that are more important to obtain a suitable predicted fundamental frequency contour.

Some preliminary work has been done using a bilingual speech corpus, Catalan/Spanish, in the touristic domain. The results are promising because the features related with the input speech shows high relevancy. This approach needs to be validated using expressive speech and languages which are more different than Catalan/Spanish.

6.3.2 Syllable and word prominence

The detection of syllable prominence is important by several aspects. The prominent syllable allows disambiguating between different meanings that differ only on the stressed syllable. Additionally syllable prominence helps to ease the understanding of the meaning of a sentence by focusing the attention of the listener in a certain word. The later can also be applied to word prominence.

Studies reveal that prominence is an important prosodic feature that affects cognitive load. Cognitive load is a measure of the degree of “mental load” that the comprehension process demands. The comprehension can be affected by several factors, such as the complexity of the text and the way it is spoken. In order to measure this cognitive load, several experiments are proposed by Delogu et al [Del98] such as listening difficulty test, comprehension and attention test.

The stress is the prosodic phenomena where a syllable is perceived as more prominent than the surrounding syllables of the word. The prominence of the syllable is correlated with some acoustical features, such as nucleus energy, midband nucleus energy (500Hz to 2000Hz), syllable duration lengthening [Cry90] and pitch variation. [Str97, Slu96, Wig94]

On the other hand, accent refers to prominence given to a syllable by the use of pitch. In this sense, accent is distinguished from the more general term stress, which is more often used to refer to all sorts of prominence (including prominence resulting from increased loudness, length or changes in sound quality), or to refer to the effort made by the speaker in producing a stressed syl lable [Roa02].

These facts are used by Tamburini [Tam04] to refer to the two components of the proposed prominence index: accent index (*en*300-2200.*dur*) and stress index (*enov.*(*Aevent.Devent*)). *Aevent* and *Devent* refer to the amplitude and duration of the Tilt events used by Tamburini to characterize the pitch contour. [Tay00].

In our work we explore the use of the acoustic parameters proposed in the literature to perform syllable and word prominence detection. The accuracy for word prominence detection is 85, 7% (*F* = 65%). This result is obtained for the speaker F1A of the Boston University Radio News Corpus [Ost95]. The experiments performed with the F1B speech corpus of Boston University Radio News Corpus show an accuracy of prominent syllable detection of 92.27%.

In our work we explore unsupervised approaches. They perform the annotation of the source waveform using indexes that merge input acoustic features. These indexes correlate its value with the strength of the existence of a given prosodic tag. Then, thresholds are used to take the decision about the presence or the absence of the prosodic label. This approach has the advantage of the simplicity: no training data is needed. However, the accuracy of such approaches is inferior than supervised approaches because of the lack of complementary sources of information that may improve the classification performance (e.g.: linguistic information).

7. References

[Ade04] Adell, J. and Bonafonte, A. Towards phone segmentation for concatenative speech synthesis. 5th ISCA Speech Synthesis Workshop, June 2004, pages 139-144, Pittsburgh, USA.

[Ade05] Adell, J., Bonafonte, A., Gómez, J. A. and Castro, M. J. Comparative study of Automatic Phone Segmentation methods for TTS , ICASSP, March 2005, Philadelphia, USA

[Agu04a] Pablo Daniel Agüero, Klaus Wimmer and Antonio Bonafonte. Automatic Analysis and Synthesis of Fujisaki's Intonation Model for TTS.. Speech Prosody 2004. Nara, Japan. March, 2004.

[Agu04b] Pablo Daniel Agüero and Antonio Bonafonte. Intonation Modeling for TTS using a Joint Extraction and Prediction Approach. 5th ISCA Speech Synthesis Workshop. Pittsburgh, EEUU . June 2004.

[Agu04c] Pablo Daniel Agüero and Antonio Bonafonte Intonation Modeling Using a joint extraction and prediction approach. 11th International Workshop "Advances in Speech Technology 2004". Maribor, Slovenia. July 2004.

[Agu04d] Pablo Daniel Agüero, Klaus Wimmer and Antonio Bonafonte. Joint Extraction and Prediction of Fujisaki's Intonation Model Parameters. ICSLP 2004. Jeju Island, Korea. October 2004.

[Baa93] R. H. Baayen, R. Piepenbrock, H. v. Rijn. 'The CELEX Lexical Database'. LDC, University of Pennsylvania, Philadelphia, USA. 1993

[Bon04] Antonio Bonafonte and Pablo Daniel Agüero. Phrase Break Prediction Using a Finite State Transducer. 11th International Workshop "Advances in Speech Technology 2004". Maribor, Slovenia. July 2004.

[Bon05] A. Bonafonte, H. Höge, H. S. Tropf, A. Moreno, H. v. d. Heuvel, D. Sündermann, U. Ziegenhain, J. Pérez, I. Kiss; Deliverable D8 of the EU project TC-STAR "Technology and corpora for Speech to Speech Translation" (FP6-506738)

[Cry90] T. H. Crystal, and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech", Journal of the Acoustical Society of America, Vol. 88, no 1, pags. 101–112, 1990.

[Del98] Cristina Delogu, Stella Conte, and Ciro Sementina, "Cognitive factors in the evaluation of synthetic speech", Speech Communication, Vol. 16, pags. 153–168, 1998.

[Dem77] A. P. Dempster, N. M. Laird, D. B. Rubin. 'Maximum Likelihood from Incomplete Data via EM Algorithm'. Journal of Royal Statistical Society. 1977

[Dux04] H. Duxans, A. Bonafonte, A. Kain, J. van Santen. 'Including Dynamic and Phonetic Information in Voice Conversion Systems'. Proc. of the ICSLP'04. Jeju Island, South Korea. 2004

[Ekl95] R. Eklund, and B. Lyberg, "Inclusion of a prosodic module in spoken language translation systems", Journal of the Acoustical Society of America, Vol. 98, no 5, pags. 2894–2899, 1995.

[Esc02] D. Escudero, and V. Cardeñoso, "Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish", Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pags. 481–484, 2002.

[Fit00] S.Fitt. 'Documentation and User Guide to Unisyn Lexicon and Post-Lexical Rules'. Technical Report. Centre for Speech Technology Research, University of Edinburgh. Edinburgh, UK, 2005

[Fuj84] H. Fujisaki, and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", Journal of the Acoustical Society of Japan, Vol. 5, pags. 233–242, 1984.

[Fuj00] H. Fujisaki, S. Ohno, and S. Narusawa, "Physiological mechanisms and biomechanical modeling of fundamental frequency control for the common japanese and the standard chinese", Proceedings of the 5th Seminar on Speech Production, pags. 145–148, 2000, bavaria, Germany.

[Gol05] C. Gollan, M.Bisani, S. Kanthak, R. Schlüter, H. Ney. 'Cross Domain Automatic Transcription on the TC-Star EPPS Corpus'. Proc. of the ICASSP'05. Philadelphia, USA, 2005

[Gom02] Gómez, J. and Castro, M.. Automatic Segmentation of Speech at the Phonetic Level. In et al., T. C., editor, Structural, Syntactic, and Statistical Pattern Recognition , volume 2396, pages 672-680.Springer-Verlag, 2002.

[Gon98] V. Goncharoff and P. Gries. 'An Algorithm for Accurately Marking Pitch Pulses in Speech Signals'. Proc. of the SIP'98. Las Vegas, USA. 1998

[Hai04] Udo Hain: ‚Phonetische Transkription für ein multilinguales Sprachsynthesesystem', PHD University of Dresden 2004

[Hof03] R. Hoffmann, O. Jokisch, D. Hirschfeld, G. Strecha. 'A Multilingual TTS System with Less Than 1 Megabyte Footprint for Embedded Applications'. Proc. of the ICASSP'03. Hong Kong, China. 2003

[Hol00] Martin Holzapfel: 'Konkatenative Speachsynthese mit großen Datenbanken', PHD University of Dresden, 2000

[Kai98] A. Kain, M. W. Macon. 'Spectral Voice Transformations for Text-to-Speech Synthesis'. Proc. of the ICASSP'98. Seattle, USA. 1998

[Kai01] A. Kain, M. W. Macon. 'Design and Evaluation of a Voice Conversion Algorithm Based on Spectral Envelope Mapping and Residual Prediction'. Proc. of the ICASSP'01. Salt Lake City, USA. 2001

[Kaw03] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, K. Shikano. 'GMM-Based Voice Conversion Applied to Emotional Speech Synthesis'. Proc. of the Eurospeech'03. Geneva, Switzerland. 2003

[Kon03] Kominek, J., Bennet, C., and Black, A. W. Evaluating and correcting phoneme segmentation for unit slection synthesis. In Proceedings of Eurospeech , pages 313-316. Geneva, Switzerland. 2003.

[Sch02] M. Schnell, O. Jokisch, R. Hoffmann, M. Küstner. 'Text-to-Speech for Low-Resource Systems'. Proc. of the MMSP'02. Virgin Islands, USA. 2002

[Ost95] M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus", 1995.

[Pri96] P. Prieto, and J. Hirschberg, "Training intonational phrasing rules automatically for English and Spanish text-to-speech", Speech Communication, Vol. 18, pags. 281–290, 1996.

[Roa02] P. Roach, "A little encyclopaedia of phonetics", University of Reading, UK, 2002.

[Slu96] A.M.C. Sluijter, and V.J. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress", Journal of the Acoustical Society of America, Vol. 100, no 4, pags. 2471–2485, 1996.

[Str97] B.M. Streefkerk, "Acoustical correlates of prominence: A design for research", Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, Vol. 21, pags. 131–142, 1997.

[Str03] G. Strecha, O. Jokisch, R. Hoffmann. 'A resource-saving modification of TD-PSOLA'. Proc. of the AST'03. Maribor, Slovenia. 2003

[Sty95] Y. Stylianou, O. Cappé, E. Moulines. 'Statistical Methods for Voice Quality Transformation'. Proc. of the Eurospeech'95. Madrid, Spain. 1995

[Sue03] D. Sündermann, H, Ney: 'synther- A new M-gram POS tagger' Proc. on Int. Conf. on Natural Language Processing and Knowledge Engineering' Beijing China, Oct 2003

[Sue04] D. Sündermann, A. Bonafonte, H. Ney, H. Höge. 'A First Step Towards Text-Independent Voice Conversion'. Proc. of the ICSLP'04. Jeju Island, South Korea. 2004

[Sue05a] D. Sündermann. 'Voice Conversion: State-of-the-Art and Future Work'. Proc. of the DAGA'05. Munich, Germany. 2005

[Sue05b] D. Sündermann, A. Bonafonte, H. Ney, H. Höge. 'A Study on Residual Prediction Techniques for Voice Conversion'. Proc. of the ICASSP'05. Philadelphia, USA. 2005

[Tam04] F. Tamburini, and C. Caini, "Automatic annotation of speech corpora for prosodic prominence", Compiling and Processing Spoken Language Corpora Workshop, pags. 53–58, 2004.

[Tay91] Taylor, P. and Isard, S. Automatic phone segmentation. In Proceedings of Eurospeech , pages 709-711. Genova, Italy, 1991.

[Tay00] P. Taylor, "Analysis and synthesis of intonation using the Tilt model", Journal of the Acoustical Society of America, Vol. 107, no 3, pags. 1697–1714, 2000.

[Tia04a] Jilei Tian, "Data-Driven Approaches for Automatic Detection of Syllable Boundaries", ICSLP 2004

[Tia04b] Jilei Tian, "Efficient Compression Method for Pronunciation Dictionaries", ICSLP 2004

[Tia04c] Jilei Tian, Jani Nurminen, " On Analysis of Eigenpitch in Mandarin Chinese", ISCSLP 2004

[Tia05] Jilei Tian, Jani Nurminen and Imre Kiss, "Optimal subset selection from text databases", ICASSP 2005

[Tol03] Toledano, D. T., Gómez, A. H., and Grande, L. V. Automatic phone segmentation. IEEE Transactions on Speech and Audio Processing , 2(6):617-625, 2003.

[Wig94] C. W. Wightman, and M. Ostendorf, "Automatic labeling of prosodic patterns", IEEE Transactions on Speech and Audio Processing, Vol. 2, no 4, pags. 469–481, 1994.

[Ye04] H. Ye, S. J. Young. 'High Quality Voice Morphing'. Proc. of the ICASSP'04. Montreal, Canada. 2004