Commercial Tools in Speech Synthesis Technology

D. Nagaraju[1], R. J. Ramasree[2], K. Kishore[3], K. Vamsi Krishna[4], R. Sujana[5]
[1]*Associate Professor, Dept. of Computer Science, Audisankara College of Engg. and Technology, Gudur, India*
[2]*Professor, Dept. of Computer Science, Rastriya Sanskrit VidyaPeet, Tirupati, India*
[3,4,5]*UG Student, Dept. of Computer Science, Audisankara College of Engg. and Technology, Gudur, India*

***Abstract*: This is a study paper planned to a new system emotional speech system for Telugu (ESST). The main objective of this paper is to map the situation of today's speech synthesis technology and to focus on potential methods for the future. Usually literature and articles in the area are focused on a single method or single synthesizer or the very limited range of the technology. In this paper the whole speech synthesis area with as many methods, techniques, applications, and products as possible is under investigation. Unfortunately, this leads to a situation where in some cases very detailed information may not be given here, but may be found in given references. The objective of the paper is to develop high quality audiovisual speech synthesis with a well synchronized talking head, primarily in Finnish. Other aspects, such as naturalness, personality, platform independence, and quality assessment are also under investigation. Most synthesizers today are so called standalones and they do not work platform independently and usually do not share common parts, thus we cannot just put together the best parts of present systems to make a state-of-the-art synthesizer. Hence, with good modularity characteristics we may achieve a synthesis system which is easier to develop and improve.**

***Keywords*: ESST, Speech Synthesis, High Quality, naturalness.**

1. Introduction

Speech is the primary means of communication between people. Speech synthesis, automatic generation of speech waveforms, has been under development for several decades (Santen et al. 1997, Kleijn et al. 1998). Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and naturalness still remain a major problem. However, the quality of present products has reached an adequate level for several applications, such as multimedia and telecommunications. With some audiovisual information or facial animation (talking head) it is possible to increase speech intelligibility considerably (Beskow et. al. 1997). Some methods for audiovisual speech have been recently introduced by for example Santen et al. (1997), Breen et al. (1996), Beskow (1996), and Le Goff et al. (1996).

The text-to-speech (TTS) synthesis procedure consists of two main phases. The first one is text analysis, where the input text is transcribed into a phonetic or some other linguistic representation, and the second one is the generation of speech waveforms, where the acoustic output is produced from this phonetic and prosodic information. These two phases are usually called as high- and low-level synthesis. The input text might be for example data from a word processor, standard ASCII from e-mail, a mobile text-message, or scanned text from a newspaper. The character string is then preprocessed and analyzed into phonetic representation which is usually a string of phonemes with some additional information for correct intonation, duration, and stress. Speech sound is finally generated with the low-level synthesizer by the information from high-level one.

The simplest way to produce synthetic speech is to play long prerecorded samples of natural speech, such as single words or sentences. This concatenation method provides high quality and naturalness, but has a limited vocabulary and usually only one voice. The method is very suitable for some announcing and information systems. However, it is quite clear that we cannot create a database of all words and common names in the world. It is maybe even inappropriate to call this speech synthesis because it contains only recordings. Thus, for unrestricted speech synthesis (text-to-speech) we have to use shorter pieces of speech signal, such as syllables, phonemes, diphones or even shorter segments.

2. Speech Production

Human speech is produced by vocal organs. The main energy source is the lungs with the diaphragm. When speaking, the air flow is forced through the glottis between the vocal cords and the larynx to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities. From the oral and nasal cavities, the air flow exits through the nose and mouth, respectively. The V-shaped opening between the vocal cords, called the glottis, is the most important sound source in the vocal system. The vocal cords may act in several different ways during speech. The most important function is to modulate the air flow by rapidly opening and closing, causing buzzing sound from which vowels and voiced consonants are produced. The fundamental frequency of vibration depends on the mass and tension and is about 110 Hz, 200 Hz, and 300 Hz with men, women, and children, respectively. With stop consonants the vocal cords may act suddenly from a completely closed position in which they cut the air flow completely, to totally open position producing a light cough or a glottal stop.

A. *Speech Synthesis METHODS*

Synthesized speech can be produced by several different methods. All of these have some benefits and deficiencies that are discussed in this and previous chapters. The methods are usually classified into three groups:

1. Articulatory synthesis, which attempts to model the human speech production system directly.
2. Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model.
3. Concatenative synthesis, which uses different length prerecorded samples derived from natural speech

B. *Speech Synthesis Tools*

Here we are introducing some of the commercial products, developing tools, and ongoing speech synthesis projects available today. It is clear that it is not possible to present all systems and products out there, but at least the most known products are presented. Some of the text in this chapter is based on information collected from Internet, fortunately, mostly from the manufacturers and developer’s official homepages. However, some criticism should be bear in mind when reading the "this is the best synthesis system ever" descriptions from these WWW-sites. First commercial speech synthesis systems were mostly hardware based and the developing process was very time-consuming and expensive. Since computers have become more and more powerful, most synthesizers today are software based systems. Software based systems are easy to configure and update, and usually they are also much less expensive than the hardware systems. However, a standalone hardware device may still be the best solution when a portable system is needed.

The speech synthesis process can be divided in high-level and low-level synthesis. A low-level synthesizer is the actual device which generates the output sound from information provided by high-level device in some format, for example in phonetic representation. A high-level synthesizer is responsible for generating the input data to the low-level device including correct text-preprocessing, pronunciation, and prosodic information. Most synthesizers contain both, high and low level system, but due to specific problems with methods, they are sometimes developed separately.

C. *Infovox*

Telia Promotor AB Infovox speech synthesizer family is perhaps one of the best known multilingual text-to-speech products available today. The first commercial version, Infovox SA-101, was developed in Sweden at the Royal Institute of Technology in 1982. The system is originally descended from OVE cascade formant synthesizer (Ljungqvist et al. 1994). Several versions of current system are available for both software and hardware platforms.

The latest full commercial version, Infovox 230, is available for American and British English, Danish, Finnish, French, German, Icelandic, Italian, Norwegian, Spanish, Swedish, and Dutch (Telia 1997). The system is based on formant synthesis and the speech is intelligible but seems to have a bit of Swedish accent. The system has five different built-in voices, including male, female, and child. The user can also create and store individual voices. Aspiration and intonation features are also adjustable. Individual articulation lexicons can be constructed for each language. For words which do not follow the pronunciation rules, such as foreign names, the system has a specific pronunciation lexicon where the user can store them. The speech rate can be varied up to 400 words per minute. The text may be synthesized also word by word or letter by letter. Also DTMF tones can be generated for telephony applications. The system is available as a half-length PC board, RS 232 connected stand-alone desktop unit, OEM board, or software for Macintosh and Windows environments (3.1, 95, NT) and requires only 486DX33MHz with 8 Mb of memory.

D. *DEC Talk*

Digital Equipment Corporation (DEC) has also long traditions with speech synthesizers. The DECtalk system is originally descended from MITalk and Klattalk. The present system is available for American English, German and Spanish and offers nine different voice personalities, four male, four female and one child. The present system has probably one of the best designed text preprocessing and pronunciation controls. The system is capable to say most proper names, e-mail and URL addresses and supports a customized pronunciation dictionary. It has also punctuation control for pauses, pitch, and stress and the voice control commands may be inserted in a text file for use by DECtalk software applications. The speaking rate is adjustable between 75 to 650 words per minute (Hallahan 1996). Also the generation of single tones and DTMF signals for telephony applications is supported.

DECtalk software is currently available for Windows 95/NT environments and for Alpha systems running Windows NT or DIGITAL UNIX. A software version for Windows requires at least Intel 486-based computer with 50 MHz processor and 8 Mb of memory. The software provides also an application programming interface (API) that is fully integrated with computer's audio subsystem. Three audio formats are supported, 16- and 8-bit PCM at 11 025 Hz sample rate for standard audio applications and 8-bit - law encoded at 8 000 Hz for telephony applications (Hallahan 1996). The software version has also three special modes, speech-to-wave mode, the log-file mode, and the text-to-memory mode. The speech-to-wave mode, where the output speech is stored into wav-file, is essential for slower Intel machines which are not able to perform real-time speech synthesis. The log-file mode writes the phonemic output in to file and the text-to-memory mode is used to store synthesized speech data into buffers from where the applications can use them (Hallahan 1996).

A hardware version of DECtalk is available as two different products, DECtalk PC2 and DECtalk Express. DECtalk PC2 is an internal ISA/EISA bus card for IBM compatible personal computers and uses a 10 kHz sample rate. DECtalk Express is an external version of the same device with standard serial interface. The device is very small (92 x194 x 33 mm, 425 g) and so suitable for portable use. DECtalk speech synthesis is also used in well-known Creative Labs Sound Blaster audio cards known as TextAssist. These have also a Voice editing tool for new voices.

E. Bell Labs Text-to-Speech

AT&T Bell Laboratories (Lucent Technologies) has also very long traditions with speech synthesis since the demonstration of VODER in 1939. The first full TTS system was demonstrated in Boston 1972 and released in 1973. It was based on articulatory model developed by Cecil Coker (Klatt 1987). The development process of the present concatenative synthesis system was started by Joseph Olive in mid-1970's (Bell Labs 1997). Present system is based on concatenation of diphones, context-sensitive allophonic units or even of triphones. The current system is available for English, French, Spanish, Italian, German, Russian, Romanian, Chinese, and Japanese (Möbius et al. 1996). Other languages are under development. The development is focused primarily for American English language with several voices, but the system is multilingual in the sense that the software is identical for all languages, except English. Some language specific information is naturally needed, which is stored externally in separate tables and parameter files. The system has also good text-analysis capabilities, as well as good word and proper name pronunciation, prosodic phrasing, accenting, segmental duration, and intonation. Bell Laboratories have particular activity for developing statistical methods for handling these problematic aspects. The latest commercial version for American English is available as several products, for example TrueTalk provided by Entropic Research and WATSON FlexTalk by AT&T.

F. Laureate

Laureate is a speech synthesis system developed during this decade at BT Laboratories (British Telecom). To achieve good platform independence Laureate is written in standard ANSI C and it has a modular architecture. The Laureate system is optimized for telephony applications so that lots of attentions have been paid for text normalization and pronunciation fields. The system supports also multi-channel capabilities and other features needed in telecommunication applications. The current version of Laureate is available only for British and American English with several different accents. Prototype versions for French and Spanish also exist and several other European languages are under development. A talking head for the system has been also recently introduced (Breen et al. 1996). More information, including several pre-generated sound examples and interactive demo, is available at the Laureate home page (BT Laboratories 1998).

G. Orator

ORATOR is a TTS system developed by Bell Communications Research (Bellcore). The synthesis is based on demisyllable concatenation (Santen 1997, Macchi et al. 1993, Spiegel 1993). The latest ORATOR version provides probably one of the most natural sounding speech available today. Special attention on text processing and pronunciation of proper names for American English is given and the system is thus suitable for telephone applications. The current version of ORATOR is available only for American English and supports several platforms, such as Windows NT, Sun, and DECstations.

H. Eurovocs

Eurovocs is a text-to-speech synthesizer developed by Technologie & Revalidatie (T&R) in Belgium. It is a small (200 x 110 x 50 mm, 600g) external device with built-in speaker and it can be connected to any system or computer which is capable to send ASCII via standard serial interface RS232. No additional software on computer is needed. Eurovocs system uses the text-to-speech technology of Lernout and Hauspie speech products described in the following chapter, and it is available for Dutch, French, German, Italian, and American English. One Eurovocs device can be programmed with two languages. The system supports also personal dictionaires. Recently introduced improved version contains also Spanish and some improvements in speech quality and device dimensions have been made.

I. Lernout & Hauspies

Lenout & Hauspies (L&H) has several TTS products with different features depending on the markets they are used. Different products are available optimized for application fields, such as computers and multimedia (TTS2000/M), telecommunications (TTS2000/T), automotive electronics (TTS3000/A), consumer electronics (TTS3000/C). All versions are available for American English and first two also for German, Dutch, Spanish, Italian, and Korean (Lernout & Hauspie 1997). Several other languages, such as Japanese, Arabic, and Chinese are under development. Products have a customizable vocabulary tool that permits the user to add special pronunciations of words which do not succeed with normal pronunciation rules. With a special transplanted prosody tool, it is possible to copy duration and intonation values from recorded speech for commonly used sentences which may be used for example in information and announcement systems. Recently, a new version for PC multimedia (TTS3000/M) has been introduced for Windows 95/NT with Software Developer's kit (API) and a special E-mail preprocessing software. The E-mail processing software is capable to interpret the initials and names in addresses and handle the header information. The new version contains also Japanese and supports run-time switching between languages. System supports wav-formats with 8 kHz and 11 kHz. The architecture is based on concatenation of rather long speech segments, such as diphones, triphones, and tetraphones.

J. Apple Plain Talk

Apple has developed three different speech synthesis systems for their MacIntosh Personal Computers. Systems have different level of quality for different requirements. The PlainTalk products are available for MacIntosh computers only and they are downloadable free from Apple homepage. MacinTalk2 is the wavetable synthesizer with ten built-in voices. It uses only 150 kilobytes of memory, but has also the lowest quality of PlainTalk family, but runs in almost every Macintosh system. MacinTalk3 is a formant synthesizer with 19 different voices and with considerably better speech quality compared to MacinTalk2. MacinTalk3 supports also singing voices and some special effects. The system requires at least Macintosh with a 68030 processor and about 300 kb of memory. MacinTalk3 has the largest set of different sounds. MacinTalkPro is the highest quality product of the family based on concatenative synthesis. The system requirements are also considerably higher than in other versions, but it has also three adjustable quality levels for slower machines. Pro version requires at least 68040 PowerPC processor with operating system 7.0 and uses about 1.5 Mb of memory. The pronunciations are derived from a dictionary of about 65,000 words and 5,000 common names.

K. Acu Voice

Acu Voice is a software based concatenative TTS system (AcuVoice 1997). It uses syllable as a basic unit to avoid modeling co-articulation effects between phonemes. Currently the system has only American English male voice, but female voice is promised to release soon. The database consists of over 60 000 speech fragments and requires about 150 Mb of hard disk space. The memory requirement is about 2.7 Mb. The system supports personal dictionaries and allows also the user to make changes to the original database. A dictionary of about 60 000 proper names is also included and names not in the dictionary are produced by letter-to-sound rules which models how humans pronounce the names which are unfamiliar to them. Additions and changes to the dictionary are also possible. The maximum speech rate is system speed dependent and is at least over 20 words per minute. The output of the synthesizer may also be stored in 8- or 16-bit PCM file. AcuVoice is available as two different products, AV1700 for standard use and AV2001 multi-channel developer's kit which is also MS-SAPI compliant. The products are available for Windows 95/NT environments with 16-bit sound card, and for Solaris x86 and SPARC UNIX workstations.

L. Cyber Talk

CyberTalk is a software based text-to-speech synthesis system for English developed by Panasonic Technologies, Inc. (PTI), USA (Panasonic 1998). The system is a hybrid formant/concatenation system which uses rule-based formant synthesis for vowels and sonorants, and prerecorded noise segments for stops and fricatives. Numbers and some alphanumerical strings are produced separately with concatenation synthesis. The CyberTalk software is available for MS Windows with male and female voices. The sound engine requires 800 kb of memory and the speech data from 360 kb to 3.5 Mb depending on voice configuration. The system has over 100,000 words built-in lexicon and separate customizable user lexicon.

M. ETI Eloquence

ETI Eloquence is a software based TTS system developed by Eloquent Technology, Inc., USA, and is currently available for British and American English, Mexican and Castillian Spanish, French, German, and Italian. Other languages, such as Chinese are also under development. For each language the system offers seven built-in voices including male, female, and child. All voices are also easily customizable by user. The system is currently available for Windows95/NT requiring at least 468 processors at 66 MHz and 8 Mb of memory, and for IBM RS/6000 workstations running AIX. Adjustable features are gender, head size, pitch baseline, pitch fluctuation, roughness, breathiness, speech, and volume. The head size is related to the vocal tract size, low pitch fluctuation produces a monotone sounding voice and a high breathiness value makes the speech sound like a whisper. The architecture consists of three main modules, the text module, the speech module, and the synthesizer. The text module has components for text normalization and parsing. The speech module uses the information from text module to determine parameter values and durations for the synthesizer. Speech is synthesized with Klattstyle synthesizer with few modifications (Hertz 1997). One special feature in the system is different text processing modes, such as math mode which converts the number 1997 as one-thousand-ninety-seven instead of nineteen ninety-seven and several spelling modes, such as radio mode which converts the input string abc as alpha, bravo, Charlie. The system also supports customized dictionaries where the user can add special words, abbreviations and roots for overriding the default pronunciation. The system can handle common difficulties with compound words, such as the th between words hothouse and mother and with common abbreviations, such as St. (saint or street).

N. Festival TTS System

The Festival TTS system was developed in CSTR at the University of Edinburgh by Alan Black and Paul Taylor and in co-operation with CHATR, Japan. The current system is available for American and British English, Spanish, and Welsh. The system is written in C++ and supports residual excited LPC and PSOLA methods and MBROLA database. With LPC method, the residuals and LPC coefficients are used as control parameters. With PSOLA or MBROLA the input may be for example standard PCM files (Black et al. 1997). As a University program the system is available free for educational, research, and individual use. The system is developed for three different aspects. For those who want simply use the system from arbitrary text-to-speech, for people who are developing language systems and wish to include synthesis output, such as different voices, specific phrasing, dialog types and so on, and for those who are developing and testing new synthesis methods. The developers of Festival are also developing speech synthesis markup languages with Bell Labs and participated development of CHATR generic speech synthesis system at ATR Interpreting Telecommunications Laboratories, Japan. The system is almost identical to Festival, but the main interests are in speech translation systems (Black et al. 1994).

O. Model Talker

ASEL ModelTalker TTS system is under development at University of Delaware, USA. It is available for English with seven different emotional voices, neutral, happy, sad, frustrated, assertive, surprise, and contradiction. English female and child voices are also under development. The system is based on concatenation of diphones and the architecture consists of seven largely independent modules, text analysis, text-to-phoneme rules, part of speech rules, prosodic analysis, discourse analysis, segmental duration calculation, and intonational contour calculation. MBROLA

The MBROLA project was initiated by the TCTS Laboratory in the Faculté Polytechnique de Mons, Belgium. The main goal of the project is to develop multilingual speech synthesis for non-commercial purposes and increase the academic research, especially in prosody generation. It is a method like PSOLA, but named MBROLA, because of PSOLA is a trademark of CNET. The MBROLA-material is available free for non-commercial and non-military purposes (Dutoit et al. 1993, 1996). The MBROLA v2.05 synthesizers is based on diphone concatenation. It takes a list of phonemes with some prosodic information (duration and pitch) as input and produces speech samples of 16 bits at the sampling frequency of the diphone database currently used, usually 16 kHz. It is therefore not a TTS system since it does not accept raw text as input, but it may be naturally used as a low level synthesizer in a TTS system. The diphone databases are currently available for American/British/Breton English, Brazilian Portuguese, Dutch, French, German, Romanian, and Spanish with male and/or female voice. Several other languages, such as Estonian, are also under development. The input data required by MBROLA contains a phoneme name, a duration in milliseconds, and a series of pitch pattern points composed of two integers each. The position of the pitch pattern point within the phoneme in percent of its total duration, and the pitch value in Hz at this position. For example, the input "_51 25 114" tells the synthesizer to produce a silence of 51 ms, and to

P. Whistler

Microsoft Whistler (Whisper Highly Intelligent Stochastic TaLkER) is a trainable speech synthesis system which is under development at Microsoft Research, Richmond, USA. The system is designed to produce synthetic speech that sounds natural and resembles the acoustic and prosodic characteristics of the original speaker and the results have been quite promising (Huang et al. 1996, Huang et al. 1997, Acero 1998). The speech engine is based on concatenative synthesis and the training procedure on Hidden Markov Models (HMM). The speech synthesis unit inventory for each individual voice is constructed automatically from unlabeled speech database using the Whisper speech recognition system (Hon et al. 1998). The use of speech recognition for labeling the speech segments is perhaps the most interesting approach for this, usually time-consuming task in concatenative synthesis. The text analysis component is derived from Lernout & Hauspie's TTS engine and, naturally, the speech engine supports MS Speech API and requires less than 3 Mb of memory.

Q. NeuroTalker

The INM (International Neural Machines, Canada) NeuroTalker is a TTS system with OCR (Optical Character Recognition) for American English with plans to release the major EU languages soon (INM 1997). The system allows the user to add specialized pronounced words and pronunciation rules to the speech database. The system recognizes most of the commonly used fonts, even when mixed or bolded. It is also capable to separate text from graphics and make corrections to text which cannot be sometimes easily corrected through an embedded speller, such as numbers or technical terms. The system requires at least Intel 486DX with 8 Mb of memory and support most scanners available. The NeuroTalker is available as two products, the standard edition with normal recognition and synthesis software, and an audiovisual edition for the visually impaired.

R. Listen 2

Listen2 is a text-to-speech system from JTS Micro Consulting Ltd., Canada, which uses the ProVoice speech synthesizer. The current system is available as an English and international version. The English version contains male and female voices and the international version also German, Spanish, French, and Italian male voices. The languages may be switched during speech and in English version the gender and pitch may be changed dynamically. The speech output may also be stored in a separate wav file. The system requires at least a 486-processor with 8 Mb of memory and a 16-bit sound card. The system has special e-mail software which can be set to announce for incoming mail with subject and sender information. The speech quality of Listen2 is far away from the best systems, but it is also very inexpensive.

S. Spruce

SPRUCE (Speech Response from UnConstrained English) is a high-level TTS system, currently under development at Universities of Bristol and Essex. The system is capable of creating parameter files suitable for driving most of the low-level synthesizers, including both formant and concatenation systems. A parallel formant synthesizer is usually used, because it gives more flexibility than other systems (Tatham et al. 1992a). In general, the system is capable to drive any low-level synthesizer based on diphones, phonemes, syllables, demi-syllables, or words (Lewis et al 1993). The system is successfully used to drive for example the DECtalk, Infovox, and CNET PSOLA synthesizers (Tatham et al. 1992b, 1995, 1996). SPRUCE architecture consists of two main modules which are written in standard C. The first on is a module for phonological tasks which alter the basic pronunciation of an individual word according to its context, and the second is a module for prosodic task which alters the fundamental frequency and duration throughout the sentence (Lewis etal. 1997). The system is based on inventory of syllables obtained from recorded natural speech to build the correct output file. The set of syllables is about 10 000 (Tatham et al. 1996). The top level of the system is dictionary based where the pronunciation of certain words are stored for several situations. For example, in weather forecast the set of used words is quite limited and consists of lots of special concepts, and with announcement systems the vocabulary may be even completely fixed. The word lexicon consists of 100 000 words which requires about 5 Mb disk space (Lewis et al. 1997).

T. Hadifix

HADIFIX (HAlbsilben, DIphone, SufFIXe) is a TTS system for German developed at University of Bonn, Germany. The system is available for both male and female voices and supports control parameters, such as duration, pitch, word prominence and rhythm. Inserting of pauses and accent markers into the input text and synthesis of singing voice are also supported. The system is based on concatenation of demisyllables, diphones, and suffixes (Portele et al. 1991, 1992). First, the input text is converted into phonemes with stress and phrasing information and then synthesized using different units. For example, the word Strolch is formed by concatenating Stro and olch. The concatenation of two segments is made by three methods. Diphone concatenation is suitable when there is some kind of stable part between segments. Hard concatenation is the simplest case of putting samples together with for example glottal stops. This also happens at each syllable boundary in demisyllable systems. Soft concatenation takes place at the segment boundaries where the transitions must be smoothed by overlapping (Portele et al. 1994). The inventory structure consists of 1080 units (750 for initial demisyllables, 150 for diphones, and 180 for suffixes) which is sufficient to synthesize nearly all German words including uncommon sound combinations originating from foreign languages (Portele et al. 1992).

U. SVOX

SVOX is a German text-to-speech synthesis system which has been developed at TIK/ETHZ (Swiss Federal Institute of Technology, Zurich). The SVOX system consists of two main modules. The transcription module includes the text analysis and the phonological generation which are speaker and voice independent. Phonological representation is generated from each input sentence and it includes the respective phoneme string, the accent level per syllable, and the phrase boundaries (position, type, and strength). The second one, phono-acoustical module, includes all the speaker dependent components that are required to generate an appropriate speech signal from the phonological representation (Pfister 1995).

V. SYNTE2 and SYNTE3

SYNTE2 was the first full text-to-speech system for Finnish and it was introduced in 1977 after five years of research in Tampere University of Technology (Karjalainen et al. 1980, Laine 1989). The system is a portable microprocessor based stand-alone device with analog formant synthesizer. The basic synthesis device consists of a Motorola 68000 microprocessor with 2048 bytes of ROM and 256 bytes of RAM, a set of special D/A-converters to generate analog control signals, and an analog signal processing part for sound generation, which is a combination of cascade and serial type formantsynthesizers. SYNTE2 takes an ASCII string as input and some special characters may be used to control features, such as speech rate, intonation, and phoneme variation (Karjalainen et al. 1980). The information hierarchy of SYNTE2 is presented in Figure 9.3. More detailed discussion of SYNTE2 see (Karjalainen 1978), (Karjalainen et al. 1980), or (Laine 1989).

3. Summary and conclusion

The product range of text-to-speech synthesizers is very wide and it is quite unreasonable to present all possible products or systems available out there. Hopefully, most of the famous and commonly used products are introduced. In near future I will introduce a new system emotional speech synthesis system for Telugu language. with 8-bit sound cards. The controllable features are the speech rate, the pitch, the pitch randomness, the peacefulness, and the duration of pauses between words. The speech rate can be adjusted between about 280 to 3200 characters per minute. The pitch can be set between 25 Hz and 300 Hz and the randomness up to 48 %. The duration of pauses between words can be set up to one second. The latest version of Mikropuhe (4.11) is available only for PC environments and it contains also singing support. All features can be also controlled by control characters within a text. The system also supports a personal abbreviation list with versatile controls and the output of the synthesizer can be stored into a separate wav-file.

References

[1] Abadjieva E., Murray I., Arnott J. (1993). Applying Analysis
[2] Laine U. (1982). PARCAS, a New Terminal Analog Model of Human Emotion Speech to Enhance Synthetic Speech. for Speech Synthesis. Proceedings of ICASSP 82 (2).
[3] Proceedings of Eurospeech 93 (2): 909-912.
[4] ModelTalker Homepage (1997). University of Delaware

[5] Acero A. (1998). Source-Filter Models for Time-Scale Pitch- (ASEL). <http://www.asel.udel.edu/speech/Dsynterf.html>.

[6] Scale Modification of Speech.Proceedings of ICASSP98. [15] Rabiner L., Shafer R. (1978). Digital Processing of Speech

[7] Belhoula K. (1993). Rule-Based Grapheme-to-Phoneme Signals, Prentice-Hall. 884. (1993). On the Use of Neural Networks in Articulatory Speech [4] Bell Laboratories TTS Homepage (1998). <http://www.bell-Synthesis. Journal of the Acoustical Society of America, JASA labs.com/project/tts/>. Campos G., Gouvea E. (1996). Speech vol. 93 (2): 1109-1121.

[8] Carlson R., Fant G., Gobl C., Granström B., Kar lsson I., Lin Q. (1989). Voice Source Rules for Text-to-Speech Synthesis. Proceedings of ICASSP 89 (1): 223-226.

[9] Carlson R., Granström B., Nord L. (1990). Evalu ation and Development of the KTH Text-to-Speech System on the Segmental Level. Proceedings of ICASSP 90 (1): 317-320.

[10] Delogu C., Paolini A., Ridolfi P., Vagges K. (1995). Intelligibility of Speech Produced by Texto-to-Speech Systems in Good and Telephonic Condtions. Acta Acoustica 3 (1995): 89-96.

[11] Dettweiler H., Hess W. (1985). Concatenation Rules for Demisyllable Speech Synthesis. Proceedings of ICASSP 85 (2): 752-755.

[12] Gonzalo E., Olaszy G., Németh G. (1993). Improvements of the Spanish Version of the MULTIVOX Text-to-Speech System. Proceedings of Eurospeech 93 (2): 869-872.

[13] HADIFIX Speech Synthesis Homepage (1997). University of Bonn. http://www.ikp.uni-bonn.de/~tpo/Hadifix.en.html

[14] Hakulinen J. (1998). Suomenkieliset puhesynteesiohjelmistot (The Software Based Speech Synthesizers for Finnish). Report Draft, University of Tampere, Department of Computing Science, Speech Interfaces, 1998.