Talking machines?! – Present and future of speech technology in Hungary

GÉZA NÉMETH, GÁBOR OLASZY, KLÁRA VICSI, TIBOR FEGYÓ

Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics

{nemeth, olaszy, vicsi, fegyo}@tmit.bme.hu

Keywords: speech technology, speech synthesis, speech recognition, dialogue systems

Speech technology has been an area of intensive research worldwide – including Hungary – for several decades. This paper will give a short overview of the challenges and results of the domain and the vision of the development and the application of the technology will also be introduced.

1. Challenges of speech technology

Speech has been the most natural and most frequently used means of human communication. Speech usually fulfills the information transmission role between biological systems *(Fig. 1)*.

The science of speech technology has emerged a few decades ago. Its' results are used in replacing certain elements of the natural speech communication chain by artificial solutions (speech recognition, speech synthesis, human-machine dialogue, diagnosis by speech, speech training, speech-to-speech translation, etc.).

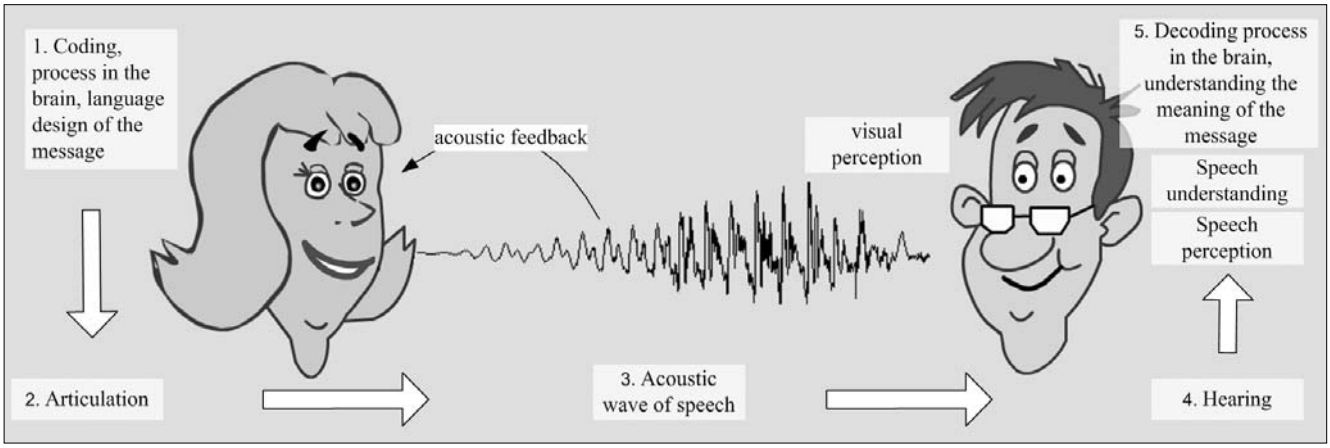
Out of the elements of the natural speech communication chain most practical engineering applications rely on the acoustic signal so in the following we shall also concentrate on this aspect. It should be noted, however, that the language is always behind the acoustic form of speech. The linguistic information determines several acoustic components of the spoken message. In order to create successful solutions of language and speech technology it is not only the processing of the acoustic signal that should be solved. Deep linguistic knowledge should also be coupled with it in order to achieve an artificial system comparable to natural communication.

The movie industry has given good visions for the practical applications of speech technology. One of the key "actors" is the HAL 9000 speaking computer in 2001: A Space Odyssey that was presented first in 1968 [26]. In 1977 in the first episode of Star Wars [1] robots perceive, store and present in many ways the multitude of information collected and transmitted through speech communication.

These visions of art created the impression for several people that all these technological breakthroughs can be reached in a short time. In practice just as interstellar spaceships, speaking and thinking robots are still to come. Because of the gap between huge expectations and significant but relatively slower technological advancements plus short time market success requirements there is a certain cyclic nature in the development of speech technology.

It is illustrated in *Fig. 2* along the dimensions of (technological) maturity and (media) visibility. The figure was created by combining Gartner's 2002 and 2006 key ICT (Information and Communication Technology) and HCI (Human Computer Interaction) forecasts in speech technology related areas. For example *natural language search* was expected to be mature for the market in 2-5 years by analysts in 2002 (i.e. between 2004-2009). The forecast

Figure 1. The process of speech communication



range changed to 5-10 years in 2006 (i.e. 2011-16). In the evaluation of *speech recognition on the desktop* similar trends can be observed. It was only *speech recognition in call centers* that moved from the 2-5 years category in 2002 into the less than two years expectation by 2006. *Text-To-Speech* (TTS) played the role of emerging technology both in 2002 and 2006 (less than two years until market penetration). It is important to note that these forecasts were created for the most developed, English speaking USA market where automation is a frequent business target (e.g. in telephone-based call centers). The real market situation varies greatly around the world, and evaluation is a continuous challenge both in Europe as a whole and in our homeland (Hungary) in particular.

In the next section of the paper an overview will be given about the results of speech technology in an international and a domestic setting with particular emphasis on existing and possible applications. In the 3rd section a short introduction will be given to the research and application vision of the area.

2. Domestic and international results of speech technology

It is worth looking at both the starting point and the current situation of R&D in domestic and international settings. It should be noted again that successful speech technology developments require advances in at least two areas: linguistic analysis and acoustic signal processing.

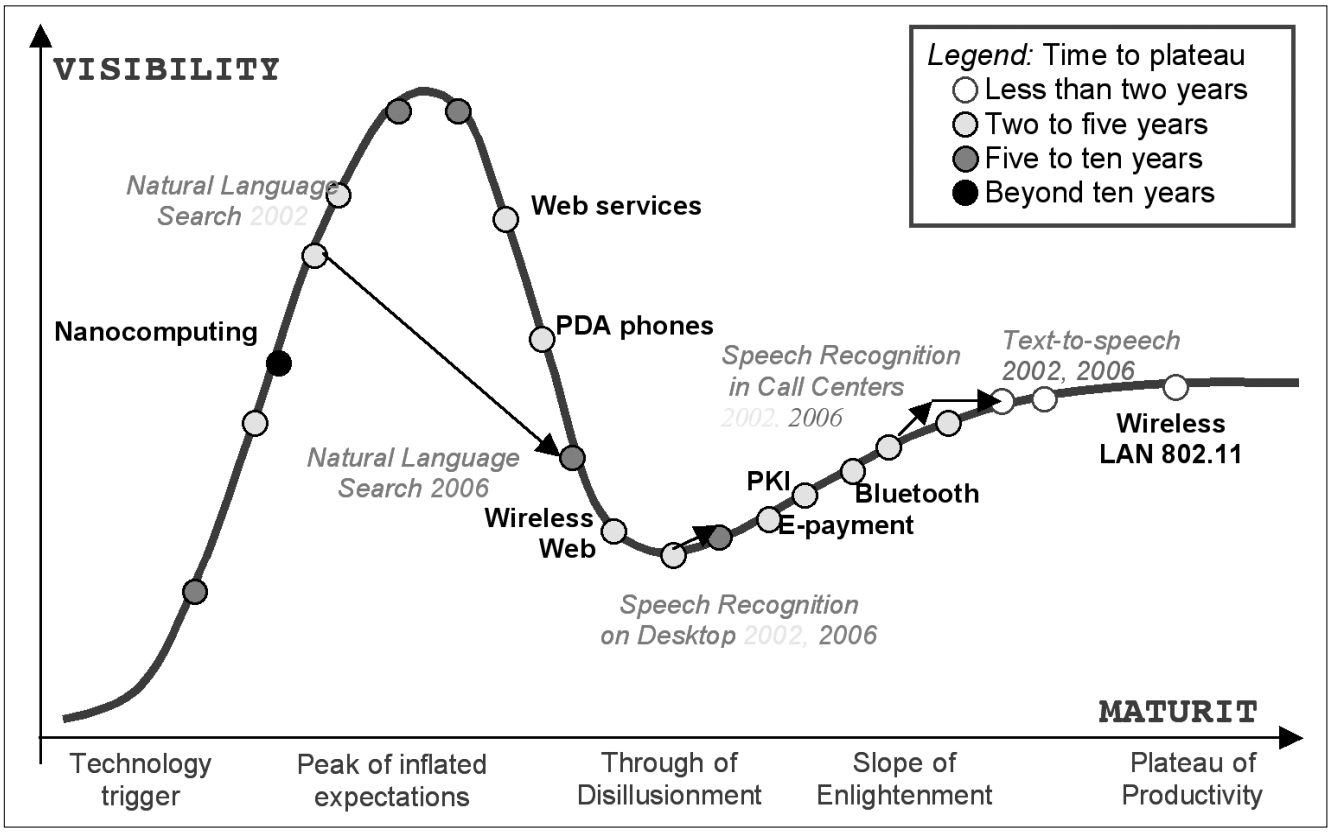
Automatic speech generation

In the area of automatic speech generation (often called speech synthesis) developers achieved as early as in 1984 that an English TTS system became part of the operating system of Apple computers [5]. This step was taken in the English versions of Microsoft Windows systems after 2000.

Hungarian research was already in the forefront of international research at the beginning of the 80s when the first general purpose, Hungarian TTS system – Hungarovox – was born in the Institute of Linguistics of the Hungarian Academy of Sciences [4]. Since then both linguistic analysis and acoustic signal processing algorithms have substantially improved. In the latter area the fourth generation is under study. The first systems modeled the human articulatory process by a time-varying filter bank and a simple excitation signal. This so-called formant synthesis solution allowed a coded acoustic database as small as a few kilobytes. The Hungarovox system was based on this technology, too. The system had a strongly robotic voice, with slow speech without rhythm and accent but with some level of intelligibility. At the predecessor of the Department of Telecommunications and Media Informatics (BME TMIT) the similar but improved MultiVox system was implemented in 12 languages [12]. The German version of MultiVox was licensed by an Austrian and a German company.

In co-operation with the developers of the Recognita optical character recognition (OCR) system we could demonstrate a Hungarian book-reading system in 1987. The first, commercially availably speech synthesizer for

Figure 2. Extended version of the Gartner Hype Cycle [Gartner Hype Cycle 2002, 2006]



the Commodore 64 computer was also developed at the same BME department [14 – p.269].

In the solutions of the second generation (from the beginning of the 90s) waveform segments were cut out from human voice, containing parts of two or three sounds (diphones and triphones, respectively). The acoustic database was compiled from these elements. The synthesized waveform was concatenated from these units (c.f. concatenative synthesis). In the following step digital signal processing algorithms were applied based on a prosodic model (pitch, timing and intensity). With this solution a speech quality resembling the given speaker could be achieved with a database on 1 to 10.000 units, and with storage space in the order of megabytes. In our research the ProfiVox system belongs to this category [13].

It was the basis in 1999 for the so-called MailMondó (MailReader) service, which read e-mail messages over the telephone for subscribers [6]. The same technology is applied in the SMS reading system for wireline telephony subscribers and in the SMSMondó (SMSReader) application for Symbian smartphones [10]. ProfiVox has also been integrated in the most widely used screen reader program for visually impaired people in Hungary (Hungarian version of Jaws).

The development of the third (corpus-based) technology started in the middle of the 90s. In this case there is no (or maybe some minor) prosodic modification by signal processing. Several hours of (usually read) speech from a speaker (or so-called voice actor/actress) are stored. This database is the acoustic database for synthesis. In case of good design there is a high probability that all sound units are available in several prosodic forms. The storage requirements of these solutions fall in the gigabyte range.

This technology is applied in the Hungarian name and address reading solution of BME TMIT that has already allowed the automation of the reverse directory service (reading out the name and the address of the subscriber based on the input phone number) of two mobile operators. In limited domains this technology can approach human quality. BME TMIT has prepared solutions for several domains. The weather forecast reader is publicly available *(www.metnet.hu)*, the automatic generation of auditory version of the price list of devices and services is applied in the IVR system of a mobile company [11]. The latest demonstration system is a railway timetable information system that “speaks” at the railway station of Sárospatak.

The fourth generation of TTS is based on Hidden Markov Models (HMMs). The basic principles are quite different from earlier TTS generations. One may say that it grew out from speech recognition experience. The acoustic basis in this case is recordings of several hours from one or more speakers (storage requirements may be in the terabyte range). These databases are used for training by statistical methods the control parameters of parametric speech coders. It is important that although a large database is required for training the resulting parameter database is typically much smaller (even just a few megabytes) which opens up several interesting applications. The quality of the latest HMM systems approach that of the third generation corpus-based systems. There are experiments with so-called hybrid systems which provide the data-driven, flexible features of HMM while maintaining the high speech quality of corpus-based systems. Our researchers conduct promising experiments in the HMM field, too [16].

It should be remembered, however, that although there are always new solutions the viability of older ones does not necessarily cease. They all have certain advantages that may be critical in certain applications. For example it is quite easy to generate whispering voice or speed up/slow down the synthetic speech with formant (or other parametric) technology which is quite difficult for corpus-based or HMM approaches.

Automatic speech recognition

The ASR has been a field of intensive research worldwide since the middle of the 20th century. From the initial sample-based systems with a vocabulary of a few words [15] the technology has advanced to large vocabulary, continuous, speaker independent technologies. The first Apple operating system containing speech recognition was announced in 1993 [5]. The latest ASR systems of industrial applications are typically based on HMMs. The basis of the technology was laid down in the 70s by the researchers of IBM. Nearly forty years have passed since then but we still cannot meet “omniscient machines” that perfectly comprehend our speech. In several narrower domains (e.g. medical dictation) though, there have been applications of regular practical use.

Current ASR systems – beyond standard software elements – have basically two language and application dependent components. Both the acoustic and the language model have to be trained according to the given application environment.

The acoustic model usually represents the speech sounds as derived from sound samples taken from several speakers. Even relatively small research databases contain at least 10 hours of speech of at least 100 speakers but there are training databases of up to several thousand hours. These samples have to be recorded in an environment that is identical (or at least similar) to the end-user application. For example there are different acoustic models for office (wideband) and telephony (narrowband) situations. The general acoustic model can be adapted to the voice of a particular speaker from a relatively smaller set of training data. The output of pattern matching based on just the acoustic models is not accurate enough. That’s why the language model of a higher level is required. It is not so surprising if we remember that human speech perception and understanding have several layers, too.

Language models help the recognizer in matching the output of the acoustic model (sound sequence) to the probable linguistic content. In fact, the individual

speech sounds are connected to a complex network according to the given application environment. In a simple case, for example command word recognition, the language model is just a simple vocabulary where the possible commands are listed. In case of the more complex continuous ASR task the linguistic probability of words following each other also have to be taken into account. In practice statistical language models trained on large text corpora are applied.

In case of agglutinative languages – such as Hungarian – because of the large number of possible word forms morpheme-based approaches have also gained momentum besides traditional word-based ones. Language models are always domain specific, the narrower the domain the higher recognition accuracy can be expected. In case of isolated command words over 95% accuracy is not rare while in case of the recognition of spontaneous conversations a result over 60% is regarded as quite good on an international scale.

Researchers of BME TMIT have participated in co-operation with industrial partners and with significant state funding in the creation of several practical applications and in the composition of related indispensable databases. Their detailed introduction is beyond the scope of the paper so only a list of major results is given below.

- *MKBF 1.0 –*
  ASR engine and development environment:
  The ASR engine is HMM-based and provides real-time processing in case of moderate size vocabularies (1000-20.000 words). The toolset supports the training of both acoustic and language models and allows N-gram models and speaker adaptation as well.
- *Medical report generator:*
  The system allows the direct speech to medical diagnosis transcription [24].
- *Classification of prosody and segmentation of speech flow:*
  Prosodic-acoustic processing speech has turned from the interest of speech synthesis research to ASR focus as well. An application – based on accent and intonation contour based classification – was developed for word and phrase boundary detection for Hungarian and Finnish. A clause segmentation and a modality detection module was also implemented [21].
- *Automatic speech-based emotion recognition* [2,17].
- *Speech databases* [22]*:*
  A large amount of labeled and annotated sound material is required for the training of ASR systems. During the preparation of databases statistical language analysis, linguistic and phonetic modeling, corpus design, database qualification and validation tasks have been completed. Diagnostic (oto-rhino-laryngology, radiology, etc.), news (Broadcast News), and audio-visual databases have also been created.
- *Multilingual speech corrector (SPECO):*
  In the framework of an EU Copernicus project a system under the fantasy name of “Magic Box” was developed. This system provides help for speech training and speech therapy audio-visually for speech- and hearing-impaired persons in Hungarian, English, German, Slovenian and Swedish. This application will be extended with a prosodic module according to our latest research results [23,20,18].
- *Keyword recognition system:*
  Recognition of a keyword without recognizing previous and following speech segments.
  Due to integrating both co-articulation and higher level pronunciation into the pattern matching process the recognition accuracy may be quite high. Because of the lack of the linguistic level this approach is not suitable for detecting short keywords. Only one keyword may be found in an announcement. The solution is definitely recommended for recognition of proper names.
- *Large vocabulary, speaker independent, real-time application optimized for the transcription of broadcast news:*
  This solution takes into account the morphology of the Hungarian language extensively. Consequently the recognition error was nearly halved compared to traditional word-based technologies. In case of speaker adaptation the word error rate was reduced below 20% on a one hour test corpus which is at state-of-the-art level compared to similar languages [7].

Although automatic speech generation and recognition technologies have improved a lot, “talking machines” (so-called *dialogue systems*) can be used only among strongly constrained situations. The reason for this is that modeling such basic phenomena of human communication as linguistic, environmental and background knowledge is still at its infancy. In case of natural dialogues we know where and to whom we are talking to and based on our earlier experience we can guess the topic of the communication, the speaking style of the speaker, etc. Most current commercial recognizers do not convey such basic information as the sex and the speaking rate of the speaker. Speech synthesizers are typically not able to change speaking styles, to express emotions and to adapt to the partner.

Speech based dialogue systems

Taking into account the above mentioned limitations there are already speech based dialogue systems operating in Hungary. Such systems currently can be successful only if the domain of the conversation is sufficiently narrow and if we inform the human user that the other partner is a machine. Such an example is the DrugLine (in Hungarian: Gyógyszervonal; *www.gyogyszervonal.hu*) [9] information system that provides web, wap and telephone based interfaces. It ensures the availability of the Patient Information Leaflets of drugs that are approved by the Hungarian National Institute of Pharmacy through three different channels. The speech based dialogue was implemented in the telephony version (adapted speech recognition and speech recognition subsystems are integrated. The phone number of the system is +36-1 8869490.

In the USA there is a widespread technology that allows the connection of an operator or an appropriate department just by pronouncing the name without keying in an extension number. A similar technology is available in Hungary as well [13], but is used by smaller organizations yet – such as some local governments – although the technology would exhibit its real advantages in case of large entities (banks, insurance companies, ministries, etc.).

3. Vision for R&D and applications of speech technology

In the field of *automatic speech generation* one of the focus areas is applying the data-driven, easier to automate HMM technology while preserving the quality of the corpus-based approach. Increasing the naturalness, social appropriateness of the generated speech is of growing importance. As a consequence, besides the general-purpose systems there are increasing numbers of outstanding quality systems in limited domains.

Besides the above mentioned solutions an important area is voice telephony access to timetables and ticket ordering of public transport systems (railways, local and long-distance buses) and providing quick access over the telephone to information of banking, insurance, state and local government systems quickly, efficiently and 7 days/24 hours.

In the area of *automatic speech recognition* efficient applications can be implemented in several domains with currently available technologies. It cannot be regarded, however, a market of out-of-the-box solutions. On the contrary, each application needs thorough preparation and pre-processing work. In order to achieve a breakthrough for more widespread applications there should be advancements in some areas.

- Noise is the hardest limit on recognition accuracy. Noise robust models and noise resistant pre-processing algorithms need even greater attention. This task is language-independent to a large extent so results for a given language may be generalized.
- Another large research area is the recognition of spontaneous conversational speech. If we look at the advancement of past decades we can recognize that technology has proceeded from well defined read speech of both acoustic and linguistic viewpoint through designed and spontaneous speech to conversational one. The last one is just as "loose" from both acoustic and linguistic viewpoint as the language of Internet forums, for example. Intensive research in this area is of great importance in order to understand natural language communication.
- In case of Hungarian and similar agglutinative languages because of the variation of word forms the size of traditional language models can easily exceed the limits of several computing platforms. More efficient modeling techniques should be defined in order to find efficient solutions for these languages (e.g. Hungarian, Finnish, Turkish, Arabic, etc.). In Hungarian the relatively free word order is another dimension of future research.

Speech technology serving public information access

Nearly half of the Hungarian population is not an Internet user. Consequently interactive information services for all the citizens can only be solved by voice-based telephony. Speech technology is the key to provide automated, cost efficient solutions to this problem. This is the only way to bridge the widening "information gap". The idea of "digital public utility" may be worth to be extended to "information public utility" (the access channel to information important for the public).

Further information can be obtained from the authors and from the Hungarian Language and Speech Technology Platform *(www.hlt-platform.hu).*

Ackowledgements

The authors acknowledge the contribution of the following speech researchers of BME TMIT – Mátyás Bartalis, András Béres, Tamás Bőhm, Tamás Csapó, Géza Gordos, Krisztián Juhász, Géza Kiss, Laczkó Klára, Péter Mihajlik, György Szaszák, Bálint Tóth, Zoltán Tüske, Ákos Viktóriusz and Csaba Zainkó – to the results presented in the paper. Research presented in the paper has been supported by among others GVOP, NKFP, Jedlik and NTP programs of the Hungarian Government.

Authors

GÉZA NÉMETH (1959) obtained his MSc in Electrical Engineering, major in Telecommunications, at the Faculty of Electrical Engineering of BME in 1983, his Dr. Univ. degree in 1987 and the PhD degree in 1997. Dr. Nemeth is the Head of the Speech Technology Laboratory of BME TMIT. His research areas include speech technology, service automation, multilingual speech and multimodal information systems, mobile user interfaces and applications.

GÁBOR OLASZY (1943) is an electrical engineer and phonetician. He graduated from the BME, Faculty of Electrical Engineering, Branch of Telecommunications, in 1967, obtained his Dr. Univ. degree – BME (1985), Ph.D. in linguistics-phonetics (1988), DSc in phonetics (2003). Research areas include acoustics of speech, development of text-to-speech systems for different languages, embedding speech synthesis into applications, research for tools towards high quality synthesised speech, combination of stored speech method with synthesised items, modelling of prosody and sound durations for TTS.

TIBOR FEGYÓ (1973) obtained his MSc in Technical Informatics at the Faculty of Electrical Engineering and Informatics of in 1997. Research areas include automatic speech recognition, acoustic and language modeling, design and development of speech information systems, speech quality measurements of telecommunications channels.

KLÁRA VICSI (1948) is a speech acoustic expert. She obtained her M.Sc. degree at Faculty of Science of Eötvös Loránd University in 1971, the Dr.Univ. degree in 1982, her Ph.D. degree in 1992 and a DSc degree from the Hungarian Academy of Sciences in 2005, She obtained a Dr. habil. title from the Budapest Univ. of Technology and Economics in 2007. Research areas inlcude speech acoustics, computer speech recognition, preparation of speech databases, psycho-acoustics, project leader in phonetics and Hungarian speech databases, providing a basis for speech recognition tasks, she was the leader of an international project of the elaboration of multi-modal speech training and development processes. She was the organizer of numerous international conferences and summer schools.

References

[1] http://en.wikipedia.org/wiki/Star_Wars_Episode_IV:_A_New_Hope

[2] European COST Action 2102 (Cross-Modal Analysis of Verbal and Nonverbal Communication), http://www.cost2102.eu/joomla/

[3] Fegyó, T., Mihajlik, P., Szarvas, M., Tatai, P., Tatai, G., "Voxenter™ – Intelligent Voice Enabled Call Center for Hungarian". In: EUROSPEECH – INTERSPEECH 2003, 8th European Conf. on Speech Communication and Technology, Geneva, Switzerland, ISCA, pp.1905–1908.

[4] Kiss Gábor, Olaszy Gábor, "Hungarovox – a Hungarian language real-time dialogue speech synthetizer system". Információ Elektronika, 2. (in Hung.), Budapest, 1984, pp.98–111.

[5] MacinTalk, http://en.wikipedia.org/wiki/PlainTalk#The_original_MacInTalk

[6] Németh G., Zainkó Cs., Fekete L., "Statistical analysis for designing and developing an e-mail reader", Híradástechnika, Vol. LVI., 2001/1, pp.23–30. (in Hung.)

[7] Mihajlik, P., Tarján, B., Tüske, Z., Fegyó, T., Investigation of Morph-based Speech Recognition Improvements across Speech Genres, Proc. of Interspeech 2009, Brighton, U.K.

[8] Németh, G., Zainkó, Cs., Kiss, G., Olaszy, G., Fekete, L., Tóth, D., Replacing a Human Agent by an Automatic Reverse Directory Service; Proc. of 15th Int. Conference on Information System Development, Budapest, Hungary, Springer LNCS, 2006, pp.323–331.

[9] Németh, G., Olaszy, G., Bartalis, M., Kiss, G., Zainkó, Cs., Mihajlik, P., Speech based Drug Information System for Aged and Visually Impaired Persons, Proc. of Interspeech 2007, Antwerp, Belgium, pp.2533–2536.

[10] Németh, G., Kiss, G., Zainkó Cs., Olaszy G., Tóth, B., Speech Generation in Mobile Phones, In: D. Gardner-Bonneau and H. Blanchard (Eds.), Human factors and interactive voice response systems, 2n edition, Springer, 2008, pp.163–191.

[11] Németh, G., Zainkó, Cs., Bartalis, M., Olaszy, G., Kiss, G., "Human Voice or Prompt Generation? Can they Co-exist in an Application?", Interspeech 2009, Brighton, UK.

[12] G. Olaszy, G. Gordos, G. Németh, The MULTIVOX multi-lingual text-to-speech converter, In: G. Bailly, C. Benoit and T. Sawallis (Eds.): Talking machines: Theories, Models and Applications, Elsevier, 1992, pp.385–411.

[13] Olaszy, G., Németh G., Olaszi, P., Kiss, G., Gordos, G., "PROFIVOX – A Hungarian Professional TTS System for Telecommunications Applications", Int. Journal of Speech Technology, Vol. 3, Nr. 3/4. Kluwer Acedemic Publ., December 2000, pp.201–216.

[14] Olaszy Gábor, Electronic speech synthesis. Műszaki Kiadó, 1989, (in Hung.).

[15] Shoebox http://www-03.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html

[16] Tóth, B., Németh, G., "Hidden markov chain-based artificial speech generation in Hungarian", Híradástechnika, 2008, pp.2–6. (in Hung.).

[17] Tóth, Sz.L., Sztahó, D., Vicsi, K., Speech Emotion Perception by Human and Machine. Proc. of COST Action 2102 International Conference, Patras, Greece, 9-31 October 2007. Revised Papers in Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, Springer, 2007, pp.213–224.

[18] Vicsi, K., Roach, P., Öster, A., Kacic, Z, Barczikay, P., Tantos, A., Csatári, F., Bakcsi, Zs, Sfakianaki, A., A multimedia, multilingual teaching and training system for children with speech disorders. International Journal of Speech Technology, Vol. 3, Kluwer Academic Publisher, 2000, pp.289–300.

[19] Vicsi K, Velkei Sz., Development of a continuous speech recognition system, 3rd Hungarian Computer Linguistics Conference, Szeged, 2005, pp.348–359., (in Hung.).

[20] Vicsi, K., Computer Assisted Pronunciation Teaching and Training Methods Based on the Dynamic Spectro-Temporal Characteristics of Speech. In: Pierre Divenyi and Georg Meyer (Eds.): "Dynamics of speech production and perception", IOS Press, Amsterdam, 2006, pp.283–307.

[21] Vicsi K, Szaszák Gy., Using Prosody for the Improvement of ASR: Sentence Modality Recognition, In: Interspeech 2008, Brisbane, Australia, 2008.

[22] http://alpha.tmit.bme.hu/speech/databases.php

[23] http://rcs.hu/sc.htm

[24] http://alpha.tmit.bme.hu/speech/research.php

[25] http://alpha.tmit.bme.hu/speech/ikta_gastro.php

[26] http://en.wikipedia.org/wiki/2001:_A_Space_Odyssey