CHALLENGES OF NATURAL LANGUAGE COMMUNICATION WITH MACHINES

DELIC, V.; SECUJSKI, M.; JAKOVLJEVIC, N.;
GNJATOVIC, M. & STANKOVIC, I.

Abstract: *The chapter gives an overview of the state of the art and the directions of future development in the field of speech technologies. The recent developments in the field of computational industry and the availability of large quantities of data have led to the emergence of applications that had been widely considered as science fiction until not long ago. Owing to these applications people can now use smart phones to send e-mail, SMS or browse the web, dictate text instead of typing it, or obtain information from interactive call centres able to e.g. detect annoyed users and allow human operators to take over. The chapter addresses several fields related to speech technology in more detail, including the basic technologies of automatic speech recognition and text-to-speech synthesis, but also specific challenges related to the role of emotion in human-machine interaction and the representation of meaning in human-machine dialogue systems. The paper concludes that, although the technologies in question are commercially available and have been widely applied, many problems remain to be solved. The available systems may be able to meet the basic needs of a user, but they are still unable to achieve the same level of communicational efficiency as a human collocutor.*

Key words: *automatic speech recognition, text-to-speech synthesis, emotions in human-machine interaction, human-machine dialogue systems, challenges for the future*

Authors´ data: Delic, V[lado]*; **Secujski**, M[ilan]*; **Jakovlevic**, N[iksa]*; **Gnjatovic**, M[ilan]*; **Stankovic**, I[gor]**, * Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia, ** European Center for Virtual Reality, Brest National Engineering School, Parvis , vdelic@uns.ac.rs; secujski@uns.ac.rs; jakovnik@uns.ac.rs, milangnjatovic@yahoo.com; bizmut@neobee.net

This Publication has to be referred as: Delic, V[lado]; Secujski, M[ilan]; Jakovlevic, N[iksa]; Gnjatovic, M[ilan] & Stankovic, I[gor] (2013) Challenges of Natural Language Communication with Machines, Chapter 19 in DAAAM International Scientific Book 2013, pp. 371-388, B. Katalinic & Z. Tekic (Eds.), Published by DAAAM International, ISBN 978-3-901509-94-0, ISSN 1726-9687, Vienna, Austria
DOI: 10.2507/daaam.scibook.2013.19

1. Introduction

As the most common way of communication between humans, speech has been considered as a convenient medium for human-machine interaction for a long time. Firstly, speech is a natural interface and humans are already fluent in it. Furthermore, while communicating with machines using speech, humans are free to simultaneously perform other tasks (Shafer, 1994). It has even been suggested that the very invention of speech by humans was not related principally to their desire to express their thoughts (for that might have been done quite satisfactorily using bodily gesture), but rather to their desire to "talk with their hands full" (Paget, 1930). Throughout the history, humans have continued to use the same communication interface not only between themselves, but also to address animals, who have been the principal technological aid to the mankind for a long time. It is therefore quite natural that, since animal power was replaced with machines, humans have been interested in the development of the technological means to extend speech communication interface to machines as well.

The design of a machine which mimics human communication capabilities in terms of understanding spoken utterances and responding to them properly has been recognized as a scientific problem for centuries (Juang & Rabiner, 2005). Since the first system for "speech analysis and synthesis" proposed in the 1930s (Dudley, 1939, Dudley et al, 1939), there has been tremendous progress in the field. The systems pretending to understand speech utterances have progressed from simple machines that respond to small sets of sounds to sophisticated systems able to respond properly to fluently spoken natural language, taking into account the statistics of the natural language in question as well as the variability introduced by different communication channels or speaker characteristics. On the other hand, the systems producing human speech have evolved from machines able to reproduce only individual speech sounds to systems able to produce sentences of natural language virtually indistinguishable from those produced by a human speaker. The research in the field of speech technology has been further accelerated by a rapid advent of powerful computing devices, leading to the emergence of a range of commercially available applications based on human-machine speech interaction, including personal assistants, dictation systems, information servers as well as aids to the disabled.

The full potential of speech as a human-machine interface can be reached only in case of natural language interfaces, which, unlike directed dialog interfaces, allow humans to communicate in the same conversational language they would use when talking to other humans (Minker & Benacef, 2004). However, the development of such an interface is burdened with the incorporation of a large quantity of domain knowledge into a very complex model of understanding, in order to be able to handle any user input successfully. Consequently, human-machine speech interaction is closely tied to the area of natural language processing (NLP), i.e. the study of computertreatment of natural human languages, including a wide variety of linguistic theories, cognitive models, andengineering approaches. With the rapid development

of the Internet, large quantities of textual and speech data have become available, which, together with the technological progress in the computer industry, enables new advances in natural language processing and the development of algorithms which may have not been computationally feasible until now. The research in the field of speech technology today focuses on a number of fields, among which the following are recognized as the most important:

- Spoken language understanding(SLU), aimed at the extraction of meaning from uttered words. When related to the conversion of a spoken utterance to a string of lexical words only, it is referred to as automatic speech recognition (ASR). Within multimodal communications systems, other input modalities including touch and image can be used together with speech (Fig. 1).
- Spoken language generation(SLG), aimed at the generation of a spoken utterance from the meaning represented according to an existing semantic model (in which case it also comprises the problem of composing a sentence that would convey particular meaning) or from a readily available string of words (in which case it is referred to text-to-speech synthesis or TTS). Within multimodal communication systems appropriate visual output can be generated as well (e.g. a talking head), enhancing the efficiency and naturalness of communication (Fig 1.).
- Human machine dialogue management, aimed at the enabling of machines to support a dialogue similar to one between humans, and related to the implementation of semantic models and dialogue processes. A dialogue strategy should be based on the language act theory, and should take into account contextual information and the knowledge of the interaction domain.

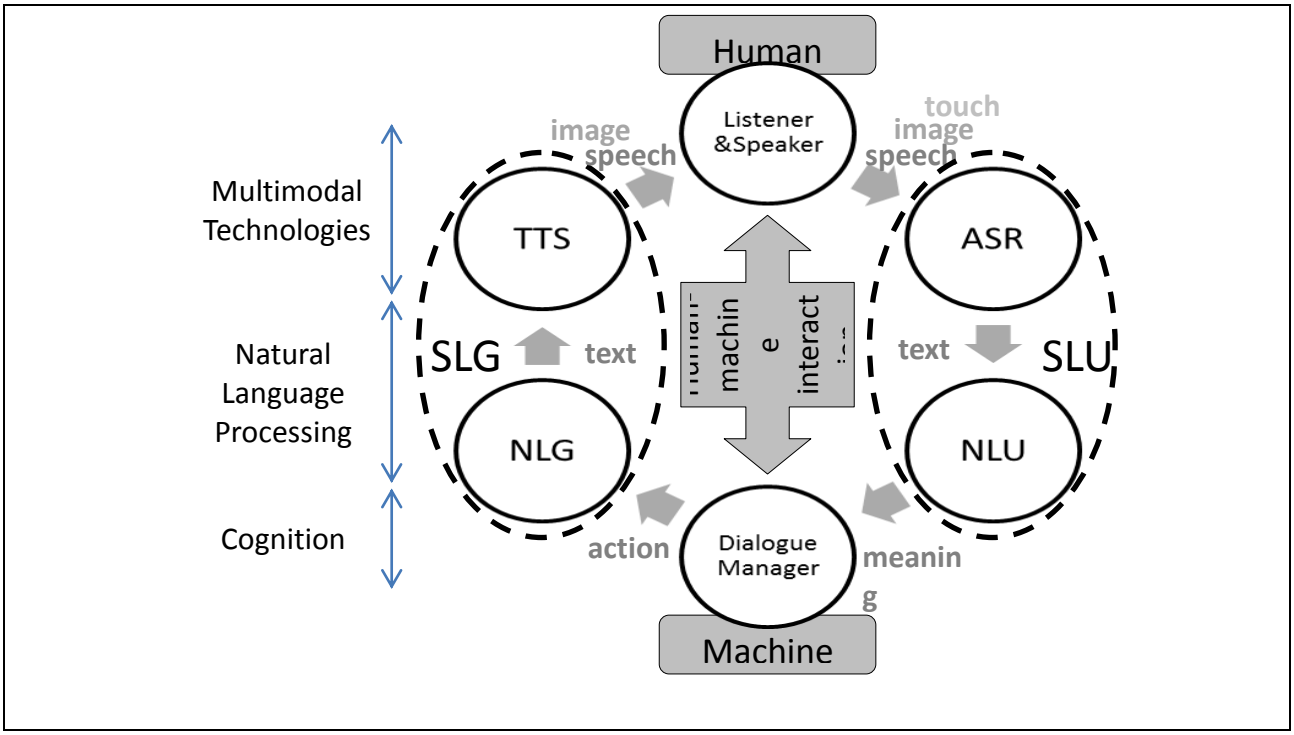

Fig. 1. Modules and modalities in human-machine interaction

- Recognition and production of vocal emotion, aimed at modelling the links between human emotion and the features of human speech communication related to it. Namely, emotion appears to be conveyed through changes in pitch, loudness, timbre, speech rate and timing which is largely independent from linguistic and semantic information.

It should be kept in mind that all of the aforementioned fields are heavily language dependent,in most cases the extent of language dependency being such that it is necessary to develop a great deal of speech and language resources and techniques independently for each language. However, models and algorithms used to treat corresponding problems across different languages are largely the same, which leads to the conclusion that there exist both global and language dependent challenges of enabling fluent human machine speech communication. Recently, great scientific and technological development has been is observed both as regards global and language specific challenges, the latter principally owing to the fact that speech and language resources have recently begun to appear for small languages as well (Delić et. al., 2010).

2. Challenges of automatic speech recognition

The term automatic speech recognition (ASR) refers to the automatic identification of the lexical content in a spoken utterance. Research in this field has been conducted for over 60 years, during which many different paradigms were explored. The early ASR systems were based on acoustic-phonetic theories which explain how phonemes are acoustically realized in spoken utterances. It was assumed that phonemes can be characterized by a set of acoustic properties that make distinction between phonemes. It was also assumed that coarticulation effects are straightforward and can be easily learned by a machine as well. In the recognition phase the first step was the segmentation of an utterance into stable acoustic regions and the assignment of possible (acoustically closest) phonemes to each segmented region, which resulted in a phoneme lattice. The second step was the determination of a valid word from the phoneme lattice, applying linguistic constrains such as a vocabulary or syntax or semantic rules (Juang & Rabiner, 2005).

In the 1970's stochastic paradigm was introduced and it became the main framework for further development of ASR in the next three decades. It was assumed that speech can be considered as a code that was transmitted over a noisy channel, and for that reason a number of algorithms from the information theory were applied with some adaptations (Jurafsky et. al., 2000). The basic premise is that hidden Markov model (HMM) state sequence can be used to describe the dynamics of phoneme sequence in a spoken utterance and the probabilistic nature of correspondence between linguistic codes and speech waveforms. Speaker and channel variability were modelled by Gaussian mixture models (GMM). In this model, the goal of the recognition process is to find the most probable HMM state

sequence, and the information theory already offered fast solutions in Viterbi and A* decoding algorithms (Huang et. al., 2001). An efficient procedure to estimate HMM parameters exists as well (Baumet. al., 1970). Additionally, the stochastic framework provides an elegant way to incorporate an acoustic model, which contains knowledge about acoustics, phonetics, channel and environment, with a language model, i.e. a system of knowledge about word constituents, order of words in a sequence etc.

There are many reasons why this statistical approach was dominant for so many years. A competitive approach was the one based on artificial neural networks (ANN), but one of the issues it faced was the temporal nature of speech signal. This can be overcome by using recurrent neural networks (Huang et. al., 2001) or hybrid models which use combinations of HMM and ANN where ANN are used to estimate HMM state emission probabilities (Bourlard & Morgan, 1993). However, in the 1990's it was challenging to set the number of parameters of ANN which can match the number of parameters of GMM because the ANN training algorithms are based on stochastic gradient descent. Additionally, the development of discriminative training algorithms for GMM based on maximum mutual information criterion and minimum phone/word error compensate forthe discriminative nature of ANN (He et. al., 2008). However, ANNs have found their role in robust feature extraction (since they can obtain class discriminative features) (Hermansky et. al., 2008) and in language modelling (since they are efficient in probability smoothing by word context similarity) (Bengio et. al., 2008).

One of the advantages of GMM-based ASR systems is the existence of efficient adaptation techniques, which transform features and/or parameters of the acoustic models to better match a given test environment. The standard adaptation technique is maximum likelihood linear regression (MLLR), which adapts Gaussian means to the test environment using maximum likelihood and reduces the word error rate significantly (Leggetter & Woodland, 1995). Nevertheless, the adaptation techniques in feature space, such as cepstral mean and variance normalization, vocal tract length normalization (Jakovljević et. al., 2009) and feature-space MLLR (Gales, 1998) in conjunction with MLLR have additionally reduced the word error rate.

The ASR systems have achieved a significant development but their performances are still 3 times worse than humans in terms of word error rate (see Fig 2.). They are far less robust to different acoustic environments (noise, reverberation, background talk, etc.), communication channels (far-field microphones, cell phones, etc.), speaker characteristics (speaking style, accents, emotional state, etc.), language characteristics (dialects, vocabulary, topic domain, etc.).

Nowadays the amount of training data for resource-rich languages is not a matter of concern, and it is thus easy to obtain more data. However, the increase in the quantity of training data cannot improve the performance of the system significantly. In the experiments the amount of data has been increased by a factor of 5, and model complexity by a factor of 6, but the relative improvement in recognition performance was only 15% (Evermann et. al., 2005). It is our belief that the next

breakthrough in ASR can be achieved by applying machine learning algorithms such as deep belief networks, graphical models and sparse representation.

2.1 Deep Belief Networks

Recently presented results suggest that further improvement in ASR can be achieved by neural-networks, more precisely,using deep belief networks (DBN) (Deng et. al., 2005). Deep belief networks are probabilistic generative models that are composed of multiple layers of stochastic latent variables (Hinton, 2009).

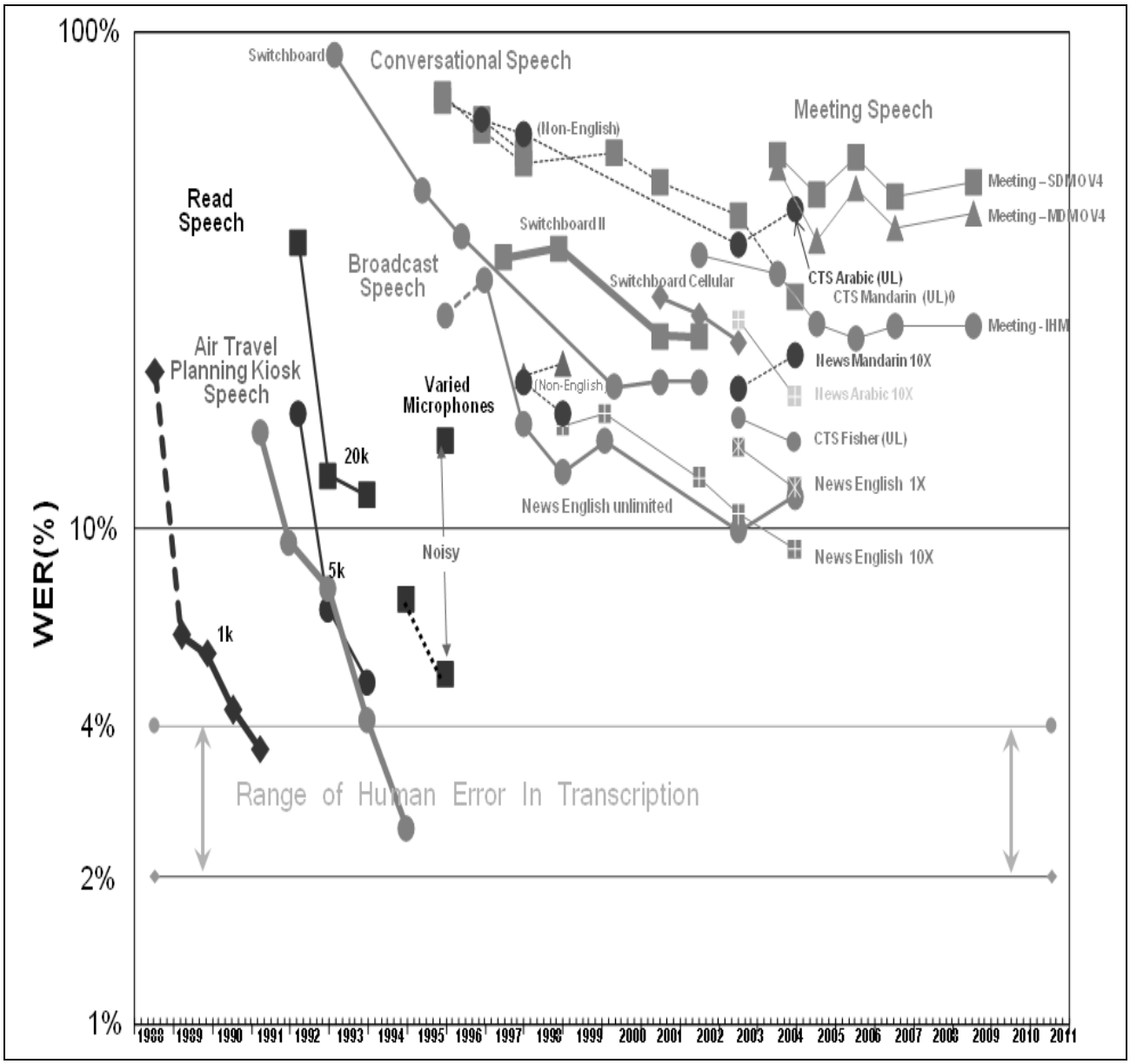

Fig. 2. Historical progress (1988-2009) on word error rates for increasingly difficult speech data(www.itl.nist.gov/iad/mig/publications/ASRhistory)

They can be viewed as multi-layer perceptrons (MLP) with many hidden layers, but the training procedure assumes learning of one layer at a time, treating the values of the latent variables of the lower layer as data for training the higher layer. The last step in the training procedure is fine-tuning by back-propagation algorithm. Such a

structure is beneficial since many simple non-linearities in each hidden layer can generate a complicated non-linearity which transforms data in a space where a linear classifier is sufficient.

It has been shown that DBN, with an appropriate training strategy and structure can result in a model which is speaker, channel and language independent (Deng et. al., 2005). The adaptation of DBN is more difficult than the adaptation of GMM, but word error rate of DBN without adaptation is far lower that the best GMM. The future work on DBN should include the pursuit for more effective deep architecture and learning algorithms.

2.2 Graphical Models

Graphical models describe dependences between a set of random variables with a graph (where each node represent random variable and edge correlation between them) (Bilmes, 2010). They provide a formal language for systematic model design and analysis, as well as efficient learning and inference techniques. Transforming an existing model into an appropriate graphical model one can get tractable and proved algorithms for learning an inference. For example restricted Boltzmann machines, which are the basis for DBN, can be transformed into Markov random fields (Deng & Li, 2013).

The standard way to obtain model-based robust ASR is to model each source of sound independently. The resulting audio file is a result of nonlinear mixing of the source models. Graphical models provide a systematic knowledge for discovering and representing such a structure and for exploiting it during inference. This approach has resulted in a multi-talker ASR algorithm which can separate and recognize the speech of four concurrently talking speakers using a single microphone (Rennie et al, 2009). It is based on tractable loopy belief propagation algorithm for iterative decoding multiple speakers. It is interesting to note that the algorithms proposed in (Rennie et al, 2009) out-performed human listeners. Similar approach can be used for speech recognition in noisy conditions as well.

2.3 Sparse Representation

The term sparse representation denotes a representation of a signal as a linear combination of a small number (smaller than the dimension of signal space) of elementary signals called atoms. The number of atoms is usually much larger than the dimension of signal space. Usually sparse representation is used for denoising, but it can be used for classification or both. The main idea is that a signal can be decomposed into atoms which model speech, and atoms which model noise, and to use only the speech part in a classification task. This idea has been tested in (Gemmeke et al, 2009) and it has outperformed the GMM model. The models which are described by examples instead by parameters are called exemplar based. These models have not been widely used in ASR community, thus they can constitute another direction for future research. Sparse representation can be used in combination with DBN (Yu et al, 2012). The results show that the model size can be

reduced significantly (by 70-90%), almost without increasingthe word error rate at all.

3. Challenges of Text-To-Speech Synthesis

The technology of text-to-speech synthesis deals with the conversion of arbitrary text into human speech in a particular language. Bridging the gap between plain text and synthesised speech with all its typical features such as intelligibility and naturalness is a complicated task, spanning multiple linguistic domains from phonetics to discourse analysis. As there is no explicit information in a plain text concerning phone durations, pitch contours nor energy variations, these factors have to be recovered from the text in the specific prosodic or expressive framework of a given speaker. The dependency of these factors from linguistic factors has to be properly modelled in order to attain high naturalness of synthesised speech (Dutoit, 1997, Morton & Tatham, 2005). The recovery of prosodic features from text is an exceedingly language dependent task referred to as high-level synthesis, and it includes natural language processing of text and its conversion into a suitable data structure describing the speech signal to be produced (referred to as the utterance plan).

The necessary steps of high-level synthesis include expanding numbers, abbreviations and other non-orthographic expressions, as well as resolution of morphological and syntactic ambiguities. A correct resolution of ambiguities is important because any error may easily lead to errors in the prosodic features of speech, impairing its naturalness. It should be kept in mind that the naturalness of synthetic speech is not merely a question of aesthetics, because incorrect intonation can mislead listeners or force them to temporarily focus their attention to lexical segmentation (identification of individual words in the input speech stream) instead of the actual meaning of the text. The largely language independent low-level synthesisis related to the production of the actual speech signal, whether by concatenation of pre-recorded segments of speech or as an output of a statistical model of speech, as in the case of hidden Markov model (HMM) based synthesis. While concatenation based techniques were the approach of choice to a majority of researchers and developers until recently, the popularity of synthesis methods based on statistical models has begun to increase, owing to their flexibility (ability to switch between speakers or speaker styles),smaller computational load and memory footprint (making them a more suitable option for environments such as portable devices), as well as speech integrity (the enhanced impression that the speech comes from a single speaker).

The focus of text-to-speech synthesis has recently shifted from intelligibility to naturalness, and some issues have not yet been addressed in a satisfactory way. Namely, current state-of-the-art speech synthesizers are still unable to produce speech which would be indistinguishable from human, and this constitutes one of the last frontiers of speech synthesis. The sources of expressiveness in human speech are not

yet well understood, and whatever they may be, they are highly variable, which further complicates the task. Furthermore, rather than addingspecific prosodic or expressive content to a "neutral" acoustic rendering of the sentence, rendering the utterance plan within a specific prosodic or expressive content thought of as a wrapper should be considered (Morton & Tatham, 2005).

The study of introducing expression into synthesized speech is related to the understanding that speech is generally influenced by intrinsic phenomena which are physical, but can nevertheless be deliberately interfered with (partially negated or enhanced), in order to "depart from listener's expectations" and thus convey particular meaning or any other feature of expressive speech. For example, sub-glottal air pressure progressively decreases with speaking, however, this does not prevent the speaker from controlling the fundamental frequency in order to convey a particular lexical accent or to give a particular word some prosodic prominence.

For that reason, the utterance specification has to be enriched with specific prosodic markup, which will reflect the changes in the expression and initiate appropriate events in the prosodic rendering of the utterance. The prosodic markup should also account for the (supposed) reaction of the speaker to the semantic or pragmatic content of the text, expressed through continuous changes in the prosodic framework including intonation, rhythm, as well as the precision of articulation. The connection between the prosodic markup and the actual prosodic rendering of the sentence is highly non-linear, and the efforts to model this relationship still fall short of producing speech which would be indistinguishable from human, principally because the gap in our understanding of how speech is produced by human beings has been underestimated.

4. Towards Emotional Speech Recognition and Synthesis

People take emotion expression and recognition for granted, but it is actually a complex process that everybody learns from the day they were born. Communication through emotions presents a huge part of everyday communication between people, and emotions are present in almost any interaction. In the near future, it will be impossible to examine any speech recognition/understanding or a speech synthesis system, or build a facial and gesture tracking system without analyzing one of the key elements of communication – emotion.

Emotions can be expressed through voice (speech emotions), face (facial expressions), and/or body (emotional body gestures). In their study, Ekman and Friesen (Ekman & Friesen,1997) discussed six emotions: happiness, sadness, surprise, anger, fear, and disgust, which became known as the "basic" emotions (Ekman & Friesen,2002), used in much related research since. Any complex emotion is considered to be a mixture of several basic ones.

The field of emotion recognition has shown tremendous potential in many areas, such as the commercial use of emotion recognition in voices in call center queuing systems (Petrushin, 2000). The use of emotion recognition technology has recently

been brought under the spotlight in terms of its potential to support countering terrorism with technology. Ball (Ball, 2011) discusses enhancing border security with automatic emotion recognition. Another possible use of emotion recognition is as an aid to speech understanding (Nicholson et al,1999). They stress that emotion in speech understanding is traditionally treated as "noise", but that a better approach would be to subtract emotions from speech and improve the performance of speech understanding systems.

A number of further applications have been proposed, which might benefit from emotion recognition components (Hone & bhadall,2004), such as intelligent tutors which change the pace or content of a computer-based tutorial based on sensing the level of interest or puzzlement of the user (Lisetti & Schiano, 2000, Picard, 1997), entertainment applications such as games or interactive movies, where the action changes based on the emotional response of the user (Nakatsu et al, 1999), help systems which detect frustration or confusion and offer appropriate user feedback (Klein, 2002) and so on. Recent studies (Stankovic et al, 2012) have shown that emotion recognition systems that utilize only speech or facial expressions do not represent a realistic way of communication and expressing emotions. In everyday life, humans use both vision and hearing to recognize emotions, thus bimodal approaches, that utilize both vision and hearing, presents a more realistic and intuitive way of detecting emotional states. It seems that people rely on both "ear" and "eye" when detecting emotions, and that for recognizing some emotions we use sound signals, while for some other emotions are more "visual" (Stankovic et al, 2012). Similarly, for expressing different emotions we employ different strategies – facial or gesture expressions, or emotional speech.

However, even though there is much work on facial expression and gesture recognition, emotion speech recognition, understanding, and synthesis, it seems that emotion recognition and synthesis is still an unsolved field. The lack of a standard, a universally agreed method for detecting and conveying emotions perhaps lies in the fact that recognizing and expressing emotions comes so naturally to us, humans, thus being particularly difficult for us to tell what distinguishes one emotion from another.

In facial expression recognition, Ekman and Friesen (Ekman & Friesen, 1977) defined facial action coding system by closely examining facial movements. Every emotion facial expression is just a combination of the movements of several facial muscles, and each basic facial movement is represented as action units (AU). Thus, presence/absence of certain AUs can tell us a lot about an expressed emotion. There are certain facial movements (AUs) that are present in almost all expressions of one emotion, and absent from expression of all other emotions.

For example, happy expression is almost always expressed by pulling lip corners – smile. However, facial expression also depends on many factors (culture, temperament, etc.), so expressions differ from one subject to another. Due to the shift of facial expression research from the recognition of acted to more spontaneous expressions, the major obstacle in the future seems to be the lack of a spontaneous expression database. Studies have proven that human behaviour becomes unnatural

the moment subjects know or suspect that they are being recorded, so it is yet to be discovered what kind of approach should be used in order to capture data with real-life spontaneous expressions in different illumination and occlusion conditions.

On the other hand, in emotional speech recognition and synthesis, we still lack a standard method for capturing and conveying emotions into speech. This is probably due to the fact that, unlike in facial expressions, there are more factors that influence speech/language, such as culture, language group, education etc., making this field of emotion recognition and synthesis a bit more challenging. While the task of emotional speech recognition is to recognize a particular emotion in human speech, the task of emotional speech synthesis is to convey it through synthesized speech. One of the main goals in this field is to detect and subtract emotional "noise" from speech, making speech recognition and comprehension easier, or synthesis a particular emotion and add it to a "flat" synthesized speech.

Many papers have studied emotions in different languages, but the future challenges would be to address multi-cultural and multi-lingual evaluations. In different languages, different speech characteristics (features) show different importance to recognize or reproduce emotions. There are several methods, such as mel-frequency cepstral coefficients (MFCC), that generally show good results in speech and emotion recognition. Also, some speech features show similar "behaviour" in most of the examined languages, and represented a starting point in any research. As Pantic and Rothkrantz (Pantic & Rothkrantz, 2003) presented, emotional speech can be examined in most of the Indo-European languages by monitoring features such as pitch, intensity, speech rate, speech contour, etc., while in some tonal languages (Thai and Vietnamese), speech emotion is more easily studied using MFCC, fundamental frequency, and zero-crossing rate (Stankovic et al, 2012).

Recently, bimodal systems have become increasingly popular, due to their stress on naturalness. These systems contain both speech emotions and facial expression, thus audio and video have their influences on one another. For example, subjects are unable to express surprise as they would express it if only expressing facial gestures without speaking. This influence of emotion speech on face movements makes it more difficult to recognize facial expressions in bimodal systems (Stankovic et al, 2012), simply because those facial expressions are less expressive, hence less informative, and more difficult to recognize. But they also represent more realistic expressions.

It is clear that in the near future more systems will focus on employing bimodal information (speech and vision), because it represents a more natural and realistic research environment, which is surely one of the main goals in the field of emotion recognition and in engineering in general. Unfortunately, bimodal systems still lack a standard database, so it is particularly difficult to compare those systems. However, with several new bimodal systems introduced every year, this problem will soon be overcome and these systems even more employed in studies.

How exactly emotional body gestures and emotions are interrelated is still an unsolved question. Most of the research is focused on detecting and synthesising

emotions in speech and facial expressions, yet our bodily movements reveal an equally significant portion of emotional state as other cues of human-human interaction. For instance, talking over phone with a person that speaks language that we do not understand is almost impossible. On the other hand, communicating with that same person face-to-face will, due to the lack of a common language, lead us to employ gestures much more in order to compensate the lack of speech. As research shows, depending on a culture, people use over 200 gestures in everyday interaction, proving that this is as an important field in emotion recognition and synthesis as face expressions and speech are.

5. Challenges of the Representation of Meaning in Dialogue Systems

The essence of the system's advanced communicative competence lies in the ability to properly interpret the user's utterance, and to adaptively manage a natural and consistent dialogue. The human-machine dialogue is natural to the extent that the system is able to address various inherently present dialogue phenomena, such as ellipses, anaphora, ungrammaticalities, meta-language, context-dependent utterances, corrections and reformulations, mixed initiative, miscommunications, uncooperative user's behavior, etc. The dialogue is consistent to the extent that the system is able to dynamically capture and represent the meaning of the dialogue, and to evaluate the user's dialogue acts with respect to it. These tasks significantly differ from the machine learning tasks such as speech recognition. At the conceptual level, they inevitably require contextual analysis, which raises the important research question of modelling the contextual information and the general knowledge of the interaction domain. In other words, an approach to meaning representation in dialogue systems should be analytically tractable and with the explanatory power.

This methodological requirement implies that the currently dominant practice of applying statistical methods to language corpora in order to derive data-driven rules for pattern recognition does not suffice. It is fair to say that the statistical approaches are quite prevalent today in the field of language processing (Chomsky, 2011), and often dogmatically anti-representational (Wilks, 2007). This trend is a consequence of relative successes of statistical approaches – in comparison with the early approaches based on logic and formal rules – over the last two decades in some aspects of machine learning. The dogma reflects in the assumption that systems may be trained to manage dialogues only by means of automated analysis of large corpora (i.e., recorded conversations).

However, the state-of-the-art in the field shows that this assumption was much too strong. Although significant scientific work was devoted and a number of sophisticated prototype systems delivered, the requirement for a natural and consistent machine dialogue still remains an elusive ideal. The general criticism levelled at statistical approaches is that they are epistemic devices taking into account only the external dialogue behaviour (cf. (Searle, 1993)), and ignoring the fact that language is a biologically innate ability that involves different linguistic and

mnemonic structures, and cognitive processes, that cannot be derived simply from language corpora (cf. Chomsky, 2011, Chomsky, 2000). We propose that taking into account the insights from behavioural and neuroimaging studies on various aspects of the human language processing system (e.g., attention, memory, etc.) is a promising research direction for further advancement of the field. The idea that the development of intelligent machines should be based on modelling people is not new (cf. (Schank, 1980)), but it is only now that results of neuroimaging studies may shape the field of human-machine interaction.

As a case in point, recent work of Gnjatović and colleagues introduces a cognitively-inspired representational approach to meaning representation in dialogue systems that integrate insights from behavioural and neuroimaging studies on working memory operations and language-impaired patients. The approach is computationally appropriate with respect that it is generalizable to different interaction domains and languages, and scalable. For detailed argument, the reader may consult (Gnjatović et al, 2012, Gnjatovic & Delic, 2012).

At the level of strategic challenges, it is reasonable to expect that the research question of adaptive dialogue management will have one of the central roles in the emerging fields of social and assistive robotics, and in the development of companion technologies. The robots' capacity to engage in a natural language dialogue would significantly – if not crucially – contribute to establishing long-term social relations to robots. Future prospects of the field of adaptive dialogue management include many challenging research problems. We shortly state some of them, although the list is by no means complete.

(1) Enabling the dialogue systems to manage multi-party dialogues in a dynamical and rich spatial context. In general, the users and the system share two interrelated contexts during the interaction – a verbal and a spatial context. Information about the spatial context is often essentially important for understanding and organizing communications (Bohus & Horovitz, 2010). It implies that the system should be aware of the surrounding environment (including the relevant interlocutors) in order to manage dialogue processes. For example, the systems should be able to robustly process linguistic inputs that instantiate different encoding patterns of motion events (e.g., bipartite and tripartite spatial scene partitioning, etc., cf. (Gnjatovic et al, 2012 and 2013)) and spatial perspectives (e.g., user-centred frame of reference, etc., cf. (Gnjatovic & Delic, 2013)).
(2) Investigation of the role of emotion and trust in human-machine interaction. An aspect of this broad research direction is focused on the investigation of linguistic cues for early recognition of negative dialogue developments, and development of dialogue strategies for preventing and handling negative dialogue development. The research on emotions is essentially supported by corpora containing samples of emotional expressions. A methodological challenge here is how to produce an appropriate, realistic emotion corpus in a

laboratory setting. Reference (Gnjatovic & Rosner, 2010) proposes a substantial refinement of the Wizard-of-Oz technique in order that a scenario designed to elicit affected behaviour in human-machine interaction could result in realistic and useful data. The proposed approach integrates two lines of research: taking into account technical requirements of a prospective spoken dialogue system, and psychological understanding of the role of the subject's motivation. The evaluation of the corpus reported in (Gnjatovic & Rosner, 2010) demonstrated that it contains recordings of overtly signalled emotional reactions whose range is indicative of the kind of emotional reactions than can be expected to occur in the interaction with spoken dialog systems. Since the subjects were not restricted by given predetermined linguistic constraints on the language to use, their utterances are indicative of the way in which non-trained, non-technical users probably like to converse with conversational agents as well.

Although in the last decade we have coincidentally witnessed the rapid increase of research interest in affected user behaviour, research in this domain is usually primarily concentrated on the detection of emotional user behaviour. However, less attention is devoted to another important research question – how to enable dialogue systems to overcome problems in the interaction related to affected user behaviour. Adaptive dialogue management is a promising research direction to address the latter question. Reference (Gnjatovic & Rosner, 2008) discusses the basic functionalities of the adaptive dialogue manager, including modelling contextual information (including the emotional state of the user), keeping track of the state of the interaction, and dynamically adapting both analytical and generative aspects of the system's behaviour according to the current state of interaction. In other words, this can be formulated as: recognizing that a problem occurred in the interaction, providing support to the user in an appropriate form – tailored to a particular problem and to the user's individual needs – and trying to advance the interaction.

(3) Enabling the dialogue systems to use reinforcement learning – e.g., by analyzing the history of interaction and the profile of the user – in order to dynamically adapt its dialogue strategy for a given user in a given situation. This is particularly important for the development of long-term collaborative conversational agents.

We take these research problems to be of great importance for increasing the level of adaptivity of human-machine dialogues. Adaptive dialogue management is of primary importance for increasing the level of naturalness of human-machine interaction and, consequently, the level of acceptance of such interfaces by users. Although considerable research effort is already to be noticed in this field, its possibilities are by no means sufficiently explored. It is to be expected that further advancements in this field will be influenced both by technological and socio-cultural trends. Currently, there is an enthusiastic and (sometimes unduly) optimistic atmosphere in the scientific community with respect to the anticipated progress. This

might by partially influenced by the fact that cognitive sciences and robotics are, at the moment, in the group of the research fields that are prioritized in fundraising activities. However, the influence is mutual – a progress in the field may have a strong influence on society, culture, economy, etc. This places even greater responsibility on the researchers. One of the certainly most difficult challenges will be to address many ethical issues raised by this technology, including military use, unethical exploitation of human social drives (Bryson, 2000), giving the users a misleading impression of the system's expertise level (Weizenbaum, 1993), etc. A great deal of the responsibility for unethical exploitations of scientific results lies with the researchers. In words of the Joseph Weizenbaum, computer scientists do not have the right to accuse politicians for leading their countries into wars – it would not be possible without computer scientists.

6. Acknowledgements

The presented study was sponsored by the Ministry of Education and Science of the Republic of Serbia under the Research grants TR32035, III44008 and OI178027. The responsibility for the content of this paper lies with the authors.

7. References

Schafer, R. W. (1994) Scientific Bases of Human-Machine Communication by Voice, in: Voice Communication Between Humans and Machines, Roe, David B., Wilpon, Jay G. (Eds.), National Academy Press, Washington D.C., USA, pp. 15–33.

Paget, R. (1930) Human Speech, Harcourt, New York, USA

Juang, B.H. & Rabiner, L. R. (2005) Automatic Speech Recognition – A Brief History of the Technology Development, in: Encyclopedia of Language and Linguistics, Brown, Keith (Ed.), Elsevier, Amsterdam, the Netherlands

Dudley, H. (1939) The Vocoder, Bell Labs Record, Bell Labs, NJ, USA,Vol. 17, pp. 122-126

Dudley, H, Riesz, R. R., Watkins, S. A. (1939) A Synthetic Speaker,J. Franklin Institute, Philadelphia, PA, USA, Vol. 227, pp. 739-764

Minker, W. & Bennacef, S. (2004) Speech and Human-Machine Dialog, Kluwer, Dordrecht, the Netherlands

Delić, V., Sečujski, M., Jakovljević, N., Janev, M., Obradović, R., Pekar, D. (2010) "Speech Technologies for Serbian and Kindred South Slavic Languages", in: Advances in Speech Recognition, Noam R. Shabtai (Ed.), SCIYO, ISBN 978-953-307-097-1, pp. 141-164

Jurafsky, D., Martin, J. H., Kehler, A., Vander Linden, K. & Ward, N. (2000) Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, Prentice-Hall Inc., Eglewood Clifs, New Jersey

Huang, X., Acero, A., & Hon, H.-W. (2001) Spoken language processing: a guide to theory, algorithm, and system development, Prentice Hall PTR New Jersey

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains The annals of mathematical statistics, JSTOR, Vol. 41, pp. 164-171

Bourlard, H. & Morgan, N. (1993) Continuous speech recognition by connectionist statistical methods Neural Networks, IEEE Transactions on, IEEE, Vol. 4, pp. 893-909

He, X., Deng, L. & Chou, W. (2008) Discriminative learning in sequential pattern recognition Signal Processing Magazine, IEEE, Vol. 25, pp. 14-36

Hermansky, H., Ellis, D. P. & Sharma, S. (2000) Tandem connectionist feature extraction for conventional HMM systems, Proc. ICASSP'00, Vol. 3, pp. 1635-1638.

Bengio, Y., Ducharme, R., Vincent, P. & Janvin, C. (2003) A neural probabilistic language model, J. Mach. Learn. Res., Vol. 3, pp. 1137-1155

Leggetter, C. J. & Woodland, P. (1995) Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models Computer Speech & Language, Elsevier, Vol. 9, pp. 171-185

Jakovljević, N., Sečujski, M. & Delić, V. (2009) Vocal Tract Length normalization strategy based on maximum likelihood criterion, Proc. of EUROCON 2009, pp. 417-420

Gales, M. J. (1998) Maximum likelihood linear transformations for HMM-based speech recognition Computer speech & language, Elsevier, Vol. 12, pp. 75-98

Evermann, G., Chan, H., Gales, M., Jia, B., Mrva, D., Woodland, P. & Yu, K. (2005) Training LVCSR systems on thousands of hours of data, Proc. of ICASSP 2005, Vol. 1, pp. 209-212

Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J. Gong, Y. & Acero, A. (2013) Recent advances in deep learning for speech research at Microsoft, Proc. of ICASSP 2013

Hinton G. E. (2009) Deep belief networks, Scholarpedia 4(5):5947

J. Bilmes, (2010) Dynamic graphical models, IEEE Signal Process. Mag.,Vol. 33, pp. 29-42

Deng, L. & Li, X. (2013) Machine learning paradigms for speech recognition: An overview, IEEE Trans. on Audio, Speech, and Lang. Process., IEEE, Vol. 21, pp. 1060–1089

Rennie, S. J., Hershey, J. R. & Olsen, P. A. (2009) Single-channel speech separation and recognition using loopy belief propagation, Proc. ICASSP 2009, pp. 3845-3848

Rennie, S. J., Hershey, J. R. & Olsen, P. A. (2009) Hierarchical variational loopy belief propagation for multi-talker speech recognition, Proc. ASRU 2009, pp. 176-181

Gemmeke, J. F., Virtanen, T. & Hurmalainen, A. (2011) Exemplar-based sparse representations for noise robust automatic speech recognition, IEEE Trans. on Audio, Speech, and Lang. Process., Vol. 19, pp. 2067-2080

Yu, D., Seide, F., Li, G. & Deng, L. (2012) Exploiting sparseness in deep neural networks for large vocabulary speech recognition, Proc. of ICASSP 2012, 4409-4412

Dutoit, T. (1997) An Introduction to Text-to-Speech Synthesis, Kluwer, Dordrecht, the Netherlands

Morton, K., Tatham, M. (2005) Developments in Speech Synthesis, Wiley, Chichester, UK

Ekman, P., Friesen, W. V. (1977) "Manual for the facial action coding system," *Consulting Psychologists Press*, Palo Alto, USA

Ekman, P., Friesen, W. V., and Hager, J. C. (2002) "Facial action coding system: a human face," Salt Lake City, Utah, USA

Petrushin, V. (2000) "Emotion recognition agents in real world," *Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium on Socially Intelligent Agents: Human in the Loop*, November 3-5, 2000, North Falmouth, Massachusetts, US

Ball, L. (2011) "Enhancing border security with automatic emotion recognition," *International Crime and Intelligence Analysis Conference 2011 (ICIAC11)*, November 2011, Manchester, UK

Nicholson, J., Takahashi, K., and Nakatsu, R. (1999) "Emotion recognition in speech using neural networks," Proceedings of the Sixth International Conference on Natural Information Processing (ICONIP'99), Perth, Australia, November 16-20, Vol. 2, pp. 495-501

Hone, K., Bhadal, A. (2004) "Affective agents to reduce user frustration: the role of agent gender," *Human-computer interaction (HCI) 2004*, Vol. 2, pp. 173-174.

Lisetti, C. L., Schiano, D. J. (2000) Automatic facial expression interpretation: where hci, artificial intelligence and cognitive science intersect, *Pragmatics and Cognition*, Vol. 8, No. 1, pp. 185-235

Picard, R. (1997) Affective Computing, The MIT Press, Cambridge, Massachusetts, USA

Nakatsu, R., Nicholson, J., and Tosa, N. (1999) Emotion recognition and its application to computer agents with spontaneous interactive capabilities, Proc. of the 7th ACM International Conference on Multimedia, October 30-November 5, 1999, Orlando, Florida, USA, pp. 343-351

Klein, J., Moon, Y., and Picard, R.W. (2002) "This computer responds to user frustration: Theory, design and results," *Interacting with Computers*, Vol. 14, pp. 119-140

Stankovic, I., Karnjanadecha, M., Delic, V. (2012) "Improvement of Thai speech emotion recognition using face feature analysis", International Review on Computers and Software (IRECOS), Vol. 7, No. 5

Ekman, P., Friesen, W. V. (1977) "Manual for the facial action coding system," *Consulting Psychologists Press*, Palo Alto, USA

Pantic, M. & Rothkrantz, L. J. M. (2003) "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, Vol. 91, No. 9, September 2003

Chomsky, N. (2011) Language and the Cognitive Science Revolution(s), Carleton University, April 8, 2011, http://chomsky.info/talks/20110408.htm

Wilks, Y. (2007) Is there progress on talking sensibly to computers? Science, Vol. 318, 927–8

Searle, J.R.(1993) The Failures of Computationalism, Think 2, 68–73

Chomsky, N. (2000) New Horizons in the Study of Language and Mind, Cambridge University Press, 2000

Schank, R.C. (1980) Language and Memory. Cognitive Science, 4:243–284

Gnjatović, M., Janev, M., Delić, V. (2012) Focus Tree: Modeling Attentional Information in Task-Oriented Human-Machine Interaction. Applied Intelligence 37(3), 305–320

Gnjatović, M., Delić, V. (2012) A Cognitively-Inspired Method for Meaning Representation in Dialogue Systems. In: Proc. of the 3rd IEEE International Conference on Cognitive Infocommunications, Kosice, Slovakia, pp. 383–388

Bohus, D., Horvitz, E.(2010) On the Challenges and Opportunities of Physically Situated Dialog. In Proc. of the AAAI Fall Symposium on Dialog with Robots, Arlington, VA, 7 pages, no pagination

Gnjatović, M., Tasevski, J., Nikolić, M., Mišković, D., Borovac, B., Delić, V. (2012) Adaptive Multimodal Interaction with Industrial Robot

In: Proc. of the IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2012), Subotica, Serbia,pp. 329–333

Gnjatović M., Tasevski J., Mišković D., Nikolić M., Borovac B., Delić V. (2013) Linguistic Encoding of Motion Events in Robotic System. In Proc. of the 6. PSU-UNS International Conference on Engineering and Technology - ICET, Novi Sad, 5 pages, no pagination

Gnjatović, M., Delić, V. (2013) Encoding of Spatial Perspectives in Human-Machine Interaction. In Proc. of the 15th International Conference SPECOM 2013, Plzeň, Czech Republic, LNAI, vol. 8113, Springer, 8 pages, in press

Gnjatović, M., Rösner, D. (2010) Inducing genuine emotions in simulated speech-based human-machine interaction: The nimitek corpus.IEEE Trans Affect Comput 1, pp. 132–144

Gnjatović, M., Rösner, D. (2008) Adaptive Dialogue Management in the NIMITEK Prototype System. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Pieraccini, R., Weber, M. (eds.) PIT 2008. LNCS (LNAI), vol. 5078, pp. 14–25. Springer, Heidelberg

Bryson, J.J. (2000) A Proposal for the Humanoid Agent-builders League (HAL). In the Proc. of The AISB 2000 Symposium on Artificial Intelligence, Ethics and (Quasi-)Human Rights, John Barnden (ed.), pp.1-6

Weizenbaum, J. (1993) Computer Power and Human Reason: From Judgement to Calculation, Penguin Books, Limited