MULTI-DIMENSIONAL MEANING ANNOTATION IN SYNTHESIS OF LISTENER VOCALIZATIONS

Viswanatha Reddy Allugunti
Glocal University UP India.

Dr. Biplab Kumar Sarkar
Glocal University UP India.
dr.bksarkar2003@yahoo.in

Dr. Raman Dugyala
Ph.D Research Co-Guide: Professor in Dept. of CSE, Vardhaman College of Engineering,
Kacharam, Shamshabad, Hyderabad, Telangana 508218, India.
raman.vsd@gmail.com

Abstract

With the ever-increasing role of computers in many areas of today's society, human- machine interaction has become an increasingly prominent part of our daily life. Ma- chines and the ways people interact with them have changed dramatically in the past few decades. Traditionally, the human-machine interfaces have often been regarded as purely rational activity, in which emotions and social aspects are secondary. This view has been changing since the mid 90's when some studies (Langer 1992; Nass and Moon 2000) demonstrate that individuals mindlessly apply social rules and expecta- tions to computers. People tend to interact with computers as if they were human-like. They unconsciously apply social rules even if they believe that such an attribution is not appropriate. Age of audience vocalizations is one of the main targets of sincerely shaded conversational speechsynthesis. Accomplishment in this undertaking relies upon the answersto three inquiries: What sorts of significance are expressed through audience vocalizations? What structure is reasonable fora given significance? Furthermore, in what setting should which listener vocalizations be delivered? In this paper, we addressthe first of these inquiries. We present a technique to record natural and expressive audience vocalizations for synthesis, and depict our way to deal with distinguish an appropriate categorical description of the significance passed on in the vocalizations. In our information, one entertainer delivers an aggregate of 967 audience vocalizations, in his normal talking style and three acted emotion-specific characters. In an open categorization scheme, we find that eleven classes happen on at minimum 5%.

Key: Multi-Dimensional, Meaning, Annotation, Synthesis, Listener, Vocalizations.

Introduction

Nowadays, humane-machine interfaces started considering emotions, social aspects and different intentions behind actual message to simulate human-like interactions. Researchers who intend to make human-like human-machine interfaces started focusing mainly on building Embodied Conversational Agents (ECAs), a kind of intelligent humanoid graphical user interfaces, which can simulate human behavior like displaying facial expressions', moving head, performing gestures and making natural interaction with others like humans do in everyday life. To maintain natural and continuous interaction with humans, ECAs have to know how to react and respond based on interpreting what humans say and their non-verbal signals. If the interaction capabilities of ECAs are to become more human-like and they are to function in social settings, their design should support continuous interaction in which all partners in an interaction perceive each other and express themselves continuously and in parallel (Thórisson 2002; Nijholt et al. 2008). In other words, human-like agents should be capable of simultaneous perception/interpretation and production of communicative behavior (Riesman et al. 2011). They should be able to signal their attitude and attention while they are listening to their interaction partner (active listening), and be able to attend to their interaction partner while they are speaking (attentive speaking). However, many ECAs still remain immobile and fall silent while listening. Active listening is a structured form of listening and responding that focuses the attention on the speaker. The essential role of listening in natural interaction – to share mutual understanding with a dialogue partner – makes it a crucial issue in the development of ECAs.

Motivation

The generation of listener vocalizations is one of the major objectives of emotion- ally colored conversational speech synthesis. It includes different research questions like where to synthesize, what to synthesize and what kind of acoustic properties to realize in order to communicate different affective states in different situations. There- fore, success in this endeavour depends on the answers to three questions: Where to synthesize a listener vocalization? What meaning should be conveyed through the synthesized vocalization? And, how to realize an appropriate listener vocalization with the intended meaning? The major motivation of this thesis is to address the latter ques- tion. The first two questions are closely linked with dialogue structure and intension planning, which are outside the scope of the present work.

Speech synthesis and interactive agents

Present day research on interactive agents has increased its focus on different spoken dialogue settings. Embodied conversational agents (ECAs) are demanding natural, spontaneous, interactive synthetic speech. Several recent investigations are aimed to reach such demands. Although the current technology met some of them such as high quality reading synthetic speech, there is a long way to travel in order to reach several objectives such as high quality interactive and spontaneous synthetic speech. This chapter provides some background information on recent work in emotional and con- versational speech synthesis. With a primary concern on listening behavior, we also discuss instructiveness in several interactive agents or virtual humans.

This chapter starts with a brief introduction on the state-of-art of speech synthe- sis technologies (see Section 3.1): unit selection approaches and statistical parametric approaches based on Hidden Markov Models (HMMs). Section 3.2 reviews some interesting investigations on spontaneous synthetic speech techniques such as expressive and conversation-like speech synthesis. We discuss the need for incorporating attentive speaking and active listening skills into ECAs. In Section 3.4, we review several ECAs, which are able to realize listening behavior, developed in the literature. Section

Speech synthesis

Speech synthesis is the process of converting text into a speech signal. The objective of Text-to-Speech (TTS) synthesis is to convert any arbitrary input text to intelligible and natural sounding speech so as to transmit information from a computer to a human. This section provides a very brief overview on the current popular speech synthesis techniques: unit-selection based and HMM-based TTS systems.

The unit-selection algorithms are well known for natural sounding speech synthe- sis. In contrast, HMM-based parametric speech synthesis is popular for intelligible systems. In addition, HMM-based speech synthesis is flexible due to its paramet- ric modeling process which can allow changing voice characteristics, emotions, and speaking styles.

Speech synthesis and interactive agents

Present day research on interactive agents has increased its focus on different spoken dialogue settings. Embodied conversational agents (ECAs) are demanding natural, spontaneous, interactive synthetic speech. Several recent investigations are aimed to reach such demands. Although the current technology met some of them such as high quality reading synthetic speech, there is a long way to travel in order to reach several objectives such as high quality interactive and spontaneous synthetic speech. This chapter provides some background information on recent work in emotional and con- versational speech synthesis. With a primary concern on listening behavior, we also discuss instructiveness in several interactive agents or virtual humans.

This chapter starts with a brief introduction on the state-of-art of speech synthe- sis technologies (see Section 3.1): unit selection approaches and statistical parametric approaches based on Hidden Markov Models (HMMs). Section 3.2 reviews some in- teresting investigations on spontaneous synthetic speech techniques such as expressive and conversation-like speech synthesis. We discuss the need for incorporating attentive speaking and active listening skills into ECAs. In Section 3.4, we review several ECAs, which are able to realize listening behavior, developed in the literature. Section

peech synthesis

Speech synthesis is the process of converting text into a speech signal. The objective of Text-to-Speech (TTS) synthesis is to convert any arbitrary input text to intelligible and natural sounding speech so as to transmit information from a computer to a human. This section provides a very brief overview on the current popular speech synthesis techniques: unit-selection based and HMM-based TTS systems.

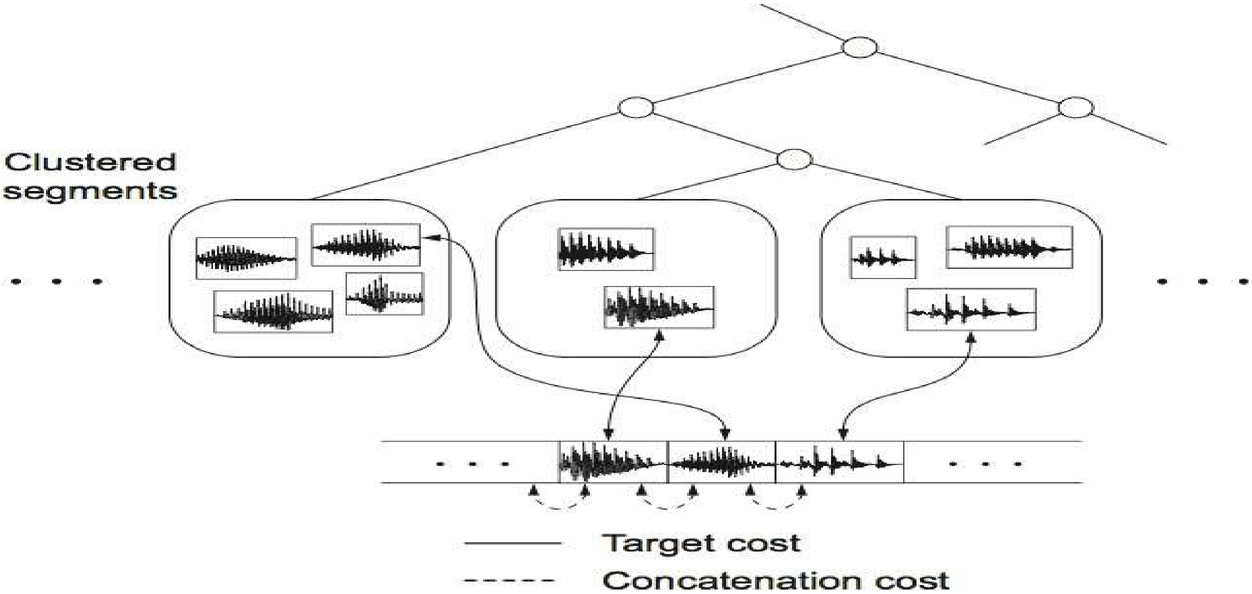

Fig.1: Multi-Dimensional Meaning Annotation in Synthesis of Listener Vocalizations

The unit-selection algorithms are well known for natural sounding speech synthe- sis. In contrast, HMM-based parametric speech synthesis is popular for intelligible systems. In addition, HMM-based speech synthesis is flexible due to its parametric modelling process which can allow changing voice characteristics, emotions, and speaking styles.

Explicit models

Explicit speech synthetic models aim for general purpose systems that are able to ex- press several emotional states based on the link between emotions and their prosodic realizations. The key for these models is to define emotion specific global prosodic settings, such as F0 level and range, speech tempo and loudness. Zovato et al. (2004) investigated signal modification techniques such as PSOLA (Pitch Synchronous and Overlap Add) to impose emotional prosody rules on selected units. This approach facilitates explicit modeling, however, it has the disadvantage of creating audible distortions for larger modifications. Schröder (2006) proposed a set of emotional prosody rules to express gradual emotions in synthetic speech based on a literature review. The hand-crafted prosody rules are implemented to reflect the prosodic settings on diphone voices available in the MARY1 text-to-speech synthesis system. The results of the work confirmed that the prosody rules are able to express a continuum of activation.

Active listening

The vocal part of the listener's behavior has been extensively discussed in Chapter 2. However, listener responses are usually multimodal in nature. Listeners use audible as well as visible acts to convey their intended meaning. Dittmann and Llewellyn (1968) confirm that most of the listener vocalizations co-occur with visual responses such as head nodes, but interestingly, many research studies on listener behavior were conducted on only one among vocal and visual modalities.

The implementation of the listening behavior has to address several research ques- tions such as when to trigger a listener response?, what to trigger? and how to realize it?. The literature indicates that, nowadays, the focus has been increased in order to investigate these research questions. This section briefly discusses some of the studies.

The open source Mary TTS platform

The current architecture of the open source MARY (Modular Architecture for Research on speech synthesis) platform is shown in Figure 4.1. MARY is a stable Java server capable of multi-threaded handling of multiple client requests in parallel. The design is

1http://mary.opendfki.de

2http://semaine.opendfki.de

Highly modular: a set of configuration files, read at system startup, define the process- ing components to use. For example, the file de.config defines the German process- ing modules, while en_US.config defines the (US) English modules. If both files are present in the configuration directory, both subsystems are loaded when starting the server. Each synthesis voice is defined by a configuration file: de-bits1.config loads the unit selection voice bits1, de-bits1-hsmm.config loads the HMM- based voice bits1-hsmm, etc. The MARY framework allows a step-by-step processing with an access to partial processing results. This framework is composed of distinct modules and has the capability of parsing speech synthesis markup such as SSML1 (Speech Synthesis Markup Language). More details on the MARY architecture can be found in Schröder and Hunecke (2007).

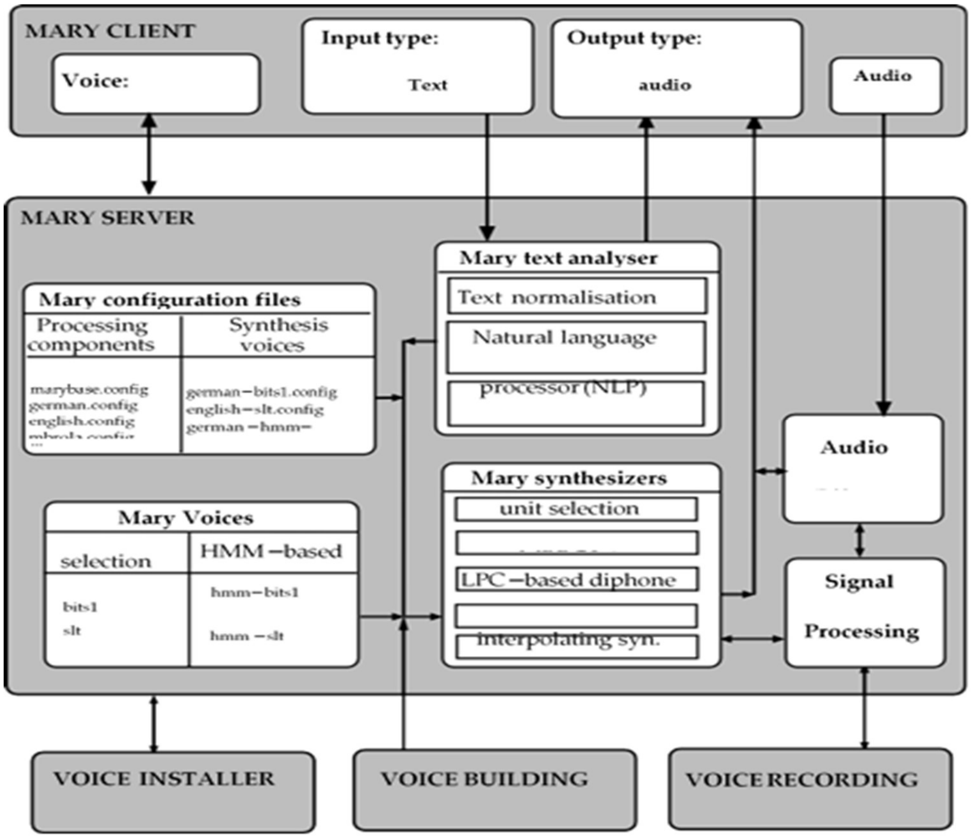

Fig.2 : Mary TTS platform version 4.0

Currently, the list of available waveform synthesizers includes a unit selection syn-thesizer (Schröder, Hunecke, and Krstulovic 2006), an MBROLA diphone synthesizer (Dutoit et al. 1996), an experimental interpolating synthesizer (Schröder 2007) and a
1http://www.w3.org/TR/speech-synthesis11
new HMM-based synthesizer ported to Java from the excellent HMM-based synthesis code from the HTS project1 (Tokuda et al. 2008). The MARY text analyzer compo- nents are described in (Schröder and Trevin 2003). The audio effects component is a new component designed to apply different effects on the audio produced by the different synthesizers. The effects are set through the audio effects GUI of the MARY client component. The Voice installer tools component is used for downloading and installing new voices or removing already installed ones. The voice recording tool is a component designed to facilitate the creation of speech synthesis databases. The voice building tools are a set components used to build new voices. The workflow of MARY framework can be divided into two stages: (i) voice build- ing process; (ii) runtime synthesis. The following sections explain these stages in de- tail.

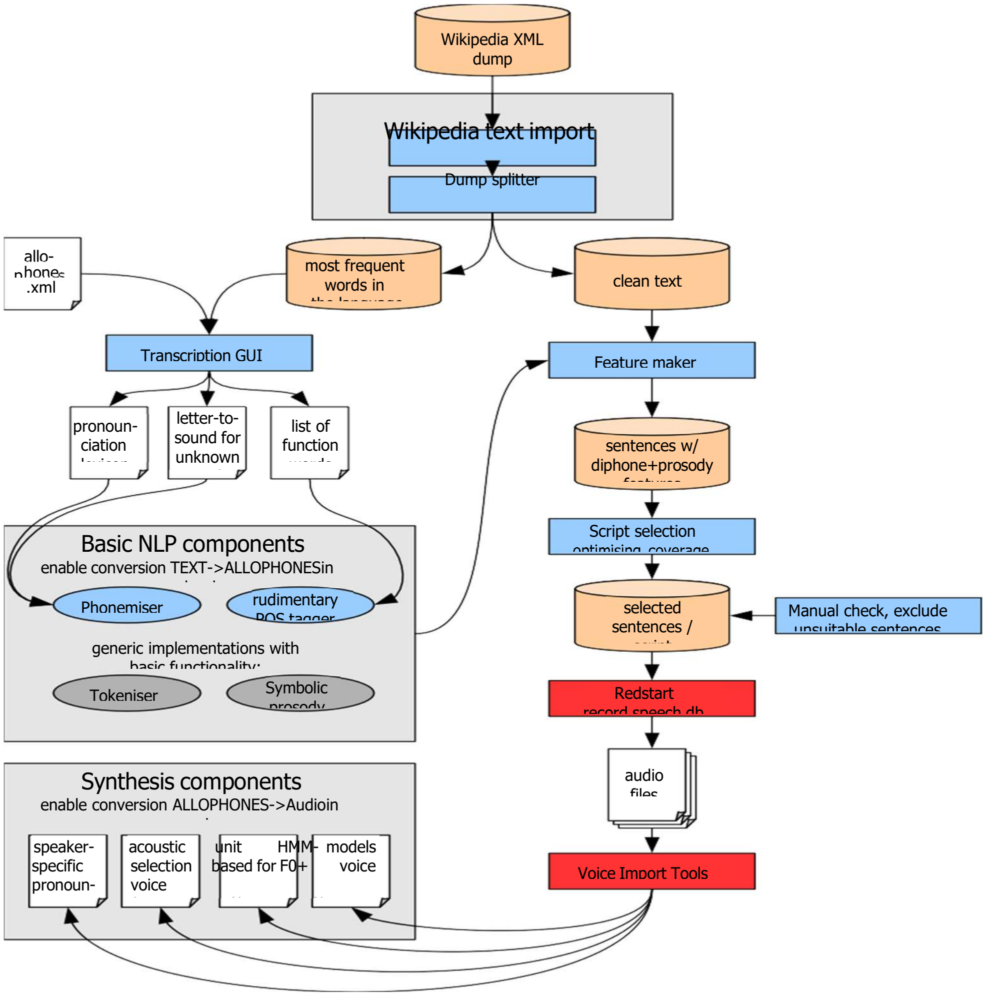

Fig.3: Multi-Dimensional Meaning Annotation in Synthesis of Listener Vocalizations Process

Whereas high-quality support of a language will usually require language-specific processing components, it is often possible to reach at least a basic support for a language using generic methods (Black and Lenz 2003). Once the NLP components have been developed, the task of creating a voice can be pursued (right branch in Figure 4.2). First, a recording script providing good diphone and prosodic coverage is selected from the text collection. Using the NLP components a feature maker component annotates each sentence in the text database with diphone and prosody features to be used in a greedy selection. The resulting collection of sentences can be used as the recording script for voice recordings with the tool Redstart. The recorded audio files can then be processed by the MARY voice import tools which generate a unit selection and/or an HMM-based voice, as well as speaker-specific prediction components for acoustic parameters. If, during the voice-building process, force-aligned transcriptions were manually corrected, it is also possible to predict speaker-specific pronunciations. In the following these steps are explained in more detail.

Methodology

The previous chapters have explained the relevant background literature required for the thesis. The rest of the chapters describe an investigation for generating listener vocalizations. This thesis is the first attempt to incorporate the ability to synthesize natural listener vocalizations in a full-scale speech synthesis system. Therefore, a systematic methodology is needed for the investigation. This chapter discusses our methodology for the investigation.

In Section 5.1, we start with identifying challenges involved in corpus-driven speech synthesis techniques to synthesize listener vocalizations. Section 5.2 list out research questions needed to address in order to achieve the objectives of this research. Section

Results

Annotators used 24 out of the 33 Baron-Cohen categories. They added nine out of the 40 categories of the emotion wheel (Scherer 2005), as well as four custom categories. The 37 categories used are shown in Table 7.3. The number of frequently used categories is much smaller, though. Only five categories were used on at least 10% of the vocalizations, and eleven categories were used on at least 5% of the data. Annotators made frequent use of the possibility to give more than one category. 17.7% of the vocalizations were labelled with a single category; 52.9% were labelled with two categories, and 29.4% with three categories.

Conclusion

In order to study each of the vocalizations per meaning, we first introduce the term meaning-vocalization combination that is used in the rest of this chapter. Each vocalization can convey maximally 11 meanings used in the corpus annotation. One stimulus indicates 11 meaning-vocalization combinations. For example, in the case of Prudence (see Table 8.4), 187 meaning-vocalization combinations (17 stimuli * 11 meaning categories) were available for analysis. Such tables for all four characters can be found in Appendix C.

Reference

1. Adel, J. (2009/2022). "Analysis and modelling of conversational elements for speech synthesis". PhD thesis. Barcelona, Spain: Universitat Politecnica de Catalunya (UPC). Ellwood, Jens, Joakim Nivre, and Elisabeth Ahlsén (1992). "On the Semantics and Pragmatics of Linguistic Feedback". In: Journal of Semantics 9.1, pp. 1–26. DOI: 10.1093/jos/9.1.1.
2. Anderson, A. and T. Lynch (1988). Listening. Oxford: Oxford University Press. Anumanchipalli, GK., K. Prahallad, and AW. Black (2011). "Festvox: Tools for Cre- Action and Analyses of Large Speech Corpora". In: Proc. Workshop on Very Large Scale Phonetics Research. UPenn, Philadelphia.
3. Atkinson, J.M. (1992). "Displaying neutrality: formal aspects of informal court proceedings". In: Talk at work: Interaction in institutional settings, pp. 199–211.
4. Baillie, J.C. (2005). "Urbi: Towards a universal robotic low-level programming language". In: Proc. International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 820–825.
5. Bales, R.F. (1950). "Interaction process analysis: A method for the study of small groups". In: Cambridge Mass.
6. Banavar, G. et al. (1999). "A case for message oriented middleware". In: Distributed Computing, pp. 846–846.
7. Baron-Cohen, S. (1988). "Social and pragmatic deficits in autism: cognitive or affective?" In: Journal of autism and developmental disorders 18.3, pp. 379–402.
8. Baron-Cohen, S. et al. (2001). "The Reading the Mind in the Eyes Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism". In: Journal of Child Psychology and Psychiatry 42.02, pp. 241–251.
9. Baron-Cohen, Simon et al. (2004). Mind Reading: The Interactive Guide to Emotions. London: Jessica Kingsley Publishers.
10. Bavelas, J.B., L. Coates, and T. Johnson (2000). "Listeners as co-narrators." In: Jour- nal of Personality and Social Psychology 79.6, p. 941.
    a. (2002). "Listener responses as a collaborative process: The role of gaze". In: Journal of Communication 52.3, pp. 566–580.
11. Bevacqua, E. (2009). "Computational model of listener behavior for Embodied Conversational Agents". PhD thesis. University Paris 8.
12. Bevacqua, E. et al. (2007). "Facial Feedback Signals for ECAs". In: AISB 2007 Annual convention, workshop "Mindful Environments". Newcastle, UK, pp. 147–153.
13. Bevacqua, E. et al. (2010). "Multimodal Backchannels for Embodied Conversational Agents". In: Proc. Intelligent Virtual Agents. Philadelphia, USA: Springer, pp. 194– 200. DOI: 10.1007/978-3-642-15892-6_21.