

Interplay between pragmatic and acoustic level to embody expressive cues in a Text to Speech system

Enrico Zovato¹, Francesca Tini-Brunozzi² and Morena Danieli¹

¹LOQUENDO S.p.A. – ²EIKON INFORMATICA

Torino, Italy

Abstract. This paper¹ deals with the problem of generating emotional speech within the Unit Selection approach to text to speech synthesis. By taking into account state-of-the-art research in different fields, from psychology to linguistics, we claim that a complex interplay between the phonetic level and the pragmatic level of language constitutes the basis of voice expression of emotions, and that the phonetic-pragmatics interplay can be accounted for in a text-to-speech system by providing accurate representations of contextually relevant discourse markers. The availability of an inventory of expressive cues implementing discourse markers, can improve the naturalness and expressivity of generated speech, moving toward the ambitious goal of emotional speech generation.

1 INTRODUCTION

In the last few years the demand of more expressive and natural human-computer interfaces has increased along with the range of applications of computer mediated communication. In particular, when natural language is the medium of the interaction between humans and computer, the complexity of the linguistic interface requires that the problem of naturalness and emotional attitude be approached at several different levels. For example, the linguistic competence of the interface should be sophisticated enough to capture the emotional state of the human conversational partner, and it should express context relevant contents in a natural-sounding, expressive, and emotional believable voice.

The fact that rhythm and intonation of the human voice are elective *loci* where emotional experiences can be represented, is considered almost an uncontroversial datum by scholars working on theoretical and experimental models of emotions within different fields, including linguistics (see [3], [7], [8], and [20]), psychology (see [9], [12] and [21]), automatic speech processing

(see [14]), neurosciences and psychoanalysis (see [18]). Of course, despite of that *consensum*, the study of how emotional states can affect voice characteristics has been approached in very different ways within those very different research areas. In the past few years, experimental psychologists focused on some individual aspects of emotional speech: In particular, their works were based on the partition between basic and complex emotions, aiming to discover their distinctive features in the acoustic signal. Most of those studies hypothesized the psychological and neurophysiological onset of a set of basic emotions, independently from other relevant aspects of the emotional experiences. For example, a huge amount of research has been devoted to the identification of the prototypical *intonation profiles* (F_0) associated with “anger”, “joy”, “pain”, “depression”, and so on (see [12], and [19] for a review). Cognitive psychologists claim that affects are the prime movers of human motivation: In this view *affects* are neurophysiologically generated, sensationlike signals. The activation of the neurobiologic affect generators triggers tone of voice as a motor stereotype, and this in turn transmits the physical data underlying empathy and emotional communication.

Despite of the tremendous advancements in the field of emotional studies, state-of-the-art research on acoustic and prosodic properties of voice expression of emotions does not provide yet an in-depth understanding of how the different features of the emotional experience show themselves in speech in terms of acoustic correlates and lexical choices. However, converging evidence from neuroscience, psychology, and linguistics can provide a promising framework for investigation. In particular, linguists take seriously into consideration the fact that speaking is a motory activity occurring on an affective basis. They showed that the analysis of speech segments can hardly show pure, prototypical intonation profiles of single emotional states if they are kept apart from the speaker’s conscious or unconscious intentions.

In our work we exploited the intention-based linguistic analysis proposed by Cresti ([7] and [8]). Cresti’s work is based on the

¹ This work has been partially supported by the EU’s 6th framework project “COMPANIONS”, contract IST 034434.

analysis of the intonation profiles of utterances from a large corpus of Italian spontaneous speech, collected in several real situations of language uses. Her results show that acoustical characteristics of emotions are influenced not only by the internal state of the speakers, but also by the external *context of occurrence* of their utterances, where the *context* includes the complex interplay between subjects' affects, and the network of interpersonal relationships where emotional experiences occur. In our view some of the difficulties in implementing emotional behaviour in artificial agents, often reported by artificial intelligence and voice scientists, reflect the importance of taking into account the way in which context affect different parameters such as lexical choices, occurrence of extra-linguistic phenomena, and the onset of voice parameters.

In this paper, we will show that useful insights in the desired direction can come from combining linguistic, acoustic and pragmatic evidences. In particular, for improving the expressiveness of the Loquendo text-to-speech system we created an inventory of discourse markers and speech acts that constitute contextually relevant part-of-speech that can be combined with neutral speech in order to generate emotionally connoted speech.

The plan of the paper is as follows: Paragraph 2 sets the problem from the point of view of text-to-speech technology, paragraph 3 proposes the linguistic analysis underlying the selection of discourse markers and speech acts, and paragraph 4 offers details related with implementation issues.

2 EXPRESSIVE SPEECH SYNTHESIS

Speech synthesis systems, also called Text to Speech systems exploit different technologies providing very different degrees of quality and naturalness. The most effective systems are the so called *corpus based synthesizers*. They are based on the concatenation of variable length speech units, selected from a large database containing speech data from a single speaker. The core technology of these systems is the search algorithm that, given the input text, has to detect the best fitting units in the database depending on the phonetic and prosodic representation of the same input.

The naturalness and intelligibility of these systems is mainly due to the fact that "exact" replicas of human speech are concatenated, avoiding any kind of signal processing that could introduce artefacts. The longer the average length of the selected units is, the better is the naturalness and acoustic quality, since fewer concatenations are necessary (every concatenation is a sort of discontinuity point). Consequently, for a given language, the goal is to design a database providing an adequate statistical coverage of the most frequent phonetic contexts. Several hours of recording sessions are therefore necessary to collect the audio data and an important point is that talents have to maintain their reading style uniform throughout the various sessions. Generally, this is a neutral "standard" reading style, i.e. no expressive attitude has to be adopted as well as no emphasis has to be introduced in the read sentences.

In this way corpus based synthesis systems, despite their intelligibility and quality, are extremely static in terms of expressive capabilities, since only one, and generally "flat" style is adopted. Adding expressivity to synthetic speech is a matter of research and investigation whose results have led to two main approaches, even if not completely satisfactory.

The first approach is based on the acquisition of speech data not only providing good phonetic coverage, but also providing a sort of expressive coverage. This is obtained by adopting different expressive styles beyond the neutral one when speech data is recorded [10]. Of course, only a limited number of styles is affordable and a preliminary choice has to be done also depending on the domain of the application. This solution is particularly effective in specific contexts, since the output quality is comparable to the one of the neutral database, but it is not flexible.

The second approach is based on the signal manipulation of the output signal obtained through the selection of speech units from a mono-stylistic (neutral) database. This kind of operation mainly aims at modifying the prosody and voice quality of the concatenated speech. In practice, the intonation path, the speech rate, intensity and spectral shape are jointly manipulated according to models that indicate how these parameters change depending on the context and the target expressive style [13,15,25]. In order to get effective models, statistical analysis of significant amounts of annotated data is necessary. The critical aspect of this approach is that, despite its flexibility, the algorithms exploited to impose the target contours often introduce distortions and compromise the naturalness of the original waveforms [23].

The paradigm here proposed is different from the two approaches previously described and less ambitious in terms of general purposes expressive synthesis. The key idea is in fact to start from a linguistic point of view that, considering the most common application domains, takes into account expressive cues that have a pragmatic function, like for example greetings, apologies, recalls, etc. These prompts are recorded and inserted into the voice database and used only in certain contexts, providing expressivity to the synthesised speech.

3 PRAGMATIC FEATURES OF THE EXPRESSIVE CUES

As discussed in the previous paragraph the goal of reaching naturalness and emotional expressivity of the synthesized speech can hardly be met by modifying only the acoustic and spectral features of the speech signal, given current state of the art of speech synthesis technologies. However, the investigation of linguistic phenomena lying at the interface between phonology and pragmatics, has showed helpful for selecting the lexical structures carrying out expressive and emotional contents.

Our goal was the creation of a rich acoustic inventory of expressive cues [5], in order to be able to integrate them in the synthesized message [11], without impairing the naturalness and the fluidity of the speech provided by the *Unit Selection* technique.

The expressive cue inventory is language specific. It includes phrases classified into different categories of speech acts [8,17] and discourse markers [3,22], and some extra-linguistic elements such as interjections, onomatopoeia and human sounds. It is worth noticing that the acoustic-prosodic and lexical structures of these phrases contribute to increase the pragmatic values of the sentences that include them. This is particularly important in a range of applications, such as human-machine interaction, e-learning, human-human computer-mediated communication, among others.

Underlying this approach is the hypothesis that the expression of emotions in human voice, can seldom, if any, be separated from the linguistic contexts [1,17] where the speech acts occur. In its turn the context affects a number of parameters of the speech act, including acoustic modifications, lexical choices, and extra-linguistic voice mediated phenomena, such as deep sigh, winks, back-channelling sounds, and so on [2,16].

Recent research in pragmatic linguistics has pointed out that the structure of human speech is modulated pragmatically [4] thanks to the competent use of discourse markers by the speakers. Bazzanella [3] offers an in-depth analysis of discourse markers, showing their potential inter-relational function. In particular, this scholar classifies discourse markers on the basis of two points of view, both of them necessary for the success of the conversation, that is the point of view of the speaker and the point of view of the co-conversant. From both point of views discourse markers are linguistics expressions that derive their meaning and pragmatic values mainly from the utterance context. For example, from the speaker's point of view, the author proposes a large set of inter-relational functions such as:

1. taking and leaving turn (i.e., *well, but, ...*)
2. fillers (i.e. *you know, see, ...*)
3. requests of attention and agreement (*can you understand this? do you agree? don't you, ...?*)
4. phatisms (*in my view, ...*)
5. request of agreement (*do you agree?...*)

From the point of view of the co-conversant, she identifies the following functions, among others:

1. interruption (*but, yes but, ...*)
2. back-channels (*aha, mhm, I see, ...*)
3. confirmation of attention (*sure, OK*)
4. phatisms (*you are welcome*)
5. reinforcement (*true, of course, ...*)

On the basis of this analysis we have identified a set of expressive cues also reported in Table 1 [24].

SPEECH ACT	EXAMPLE
Refuse	<i>Absolutely not! ...</i>
Approval	<i>Exact! ...</i>
Disapproval	<i>Absurd! ...</i>
Recall	<i>Let's keep in touch! ...</i>
Announce	<i>Here I am! ...</i>
Request of Confirmation	<i>Isn't it? ...</i>
Request of Information	<i>Why? ...</i>
Request of Action	<i>Help! ...</i>
Prohibition	<i>This is forbidden! ...</i>
Contrast	<i>I don't think so! ...</i>
Disbelief	<i>That's unbelievable! ...</i>
Surprise	<i>What a surprise! ...</i>
Regret	<i>I'm so sorry! ...</i>
Thanks	<i>Thanks a lot!</i>
Greetings	<i>Welcome! ...</i>
Apologies	<i>I'm sorry! ...</i>
Compliments	<i>Congratulations! ...</i>

Table 1. Speech acts categories with examples

The items and phrases we selected are representative of speech acts that reflect the speaker's attitude with respect to her/his conversational partner in different contexts of uses. For doing this, the linguistic analyses have been done on a corpus basis.

Also the communicative potential of human sounds is relevant. For example, a throat could communicate embarrassment, distancing, request of attention, or it could play the role of *back-channelling*. The inventory also includes human sounds as throats, bitter and hearty laughs, sobbings.

On the basis of this inventory, we could implement the acoustic counterpart of a limited, but rich, set of speech acts, including: refuse, approval/disapproval, recall in proximity, announce, request of information, request of confirmation, request of action/behaviour, prohibition, contrast, disbelief, surprise/astonishment, regret, thanks, greetings, apologies, and compliments.

4 IMPLEMENTATION OF THE EXPRESSIVE FEATURES

The design of the speech acts corpus, as previously explained, is based on linguistic rather than phonetic criteria. There is a substantial difference in the way these data are recorded with respect to the acquisition of the baseline voice data, where emphasis and too marked intonation movements are avoided. In fact, in this case, we asked our speaker to adopt the more suitable voice registry according to the semantic and pragmatic function of the scripts. Nevertheless, during the recording sessions, talents had to be accurately directed, particularly concerning the level of activation. This hasn't to be too high because the stylistic difference with the base synthetic voice would be too marked and consequently judged as unnatural. The main goal is adding expressivity without compromising the continuity of prosodic patterns. As concerns the acquisition of the speech data, for each linguistic category a set of samples is recorded. Some sets are bigger than others depending on the variety of the speech acts and on their frequency in the considered spoken language. Generally, for each voice, the database is composed of about 500 expressive utterances. The speaker is free to interpret the scripts and adopt the suitable attitude while the director only controls his/her level of activation and the pronunciation accuracy. At the end of the acquisition the best samples are selected in terms of acoustic quality, effectiveness and reliability. These data are then normalised to better match the acoustic characteristics of the base voice data and the same coding is also applied. The expressive speech data corresponding to the illocutionary acts is also automatically labelled like the neutral speech data. In fact phonetic and prosodic labels are assigned to each elementary unit (phoneme). One more label identifies the stylistic class of the utterance which the unit was extracted from. These classes are, for example, declarative, interrogative, marked interrogative, exclamation, etc.

In the selection phase the TTS avoids mixing units belonging to different categories and, in particular, will choose the expressive utterances only when an exact matching with the phonetic counterpart of the graphemic input string and the target stylistic class occur. Regarding the latter this is simply obtained through the analysis of the final punctuation of the sentence. On the contrary, if the marked input text has no correspondence in

the expressive set of data, then the concatenation of neutral speech segments is exploited.

Beyond the expressive utterances, the paralinguistic events are also recorded (laughs, coughs, hesitations, etc.). Of course, these data are not analysed at segmental and phonetic level. Only one identification label is assigned to each of them as a whole. The “synthesis” of these events is realised by inserting in the input text these labels preceded by a special control tag.

In order to make the expressive features effective, they have to be a priori known by the user. To this end we have developed a client application suitable for producing vocal prompts. One important feature of this application is the possibility to show all the available linguistic and paralinguistic events, having them classified according to the previously described categories. In this way the user can easily choose and insert the expressive cues into the synthesised speech, obtaining a more colourful synthetic speech, in terms of intonation movements and voice quality.

5 CONCLUSIONS

In this paper we approached the problem of generating emotional speech within the Unit Selection approach to text to speech synthesis by taking advantage of state-of-the-art research in different fields, from psychology to linguistics. A common evidence of research in those areas concerns the relevance of the utterance contexts, and the identification of different levels from where each human speech act is marked in terms of intentionality, expressivity, content, and emotion expressions. These results are in some sense disruptive for the traditional organization of levels of linguistic models. Actually, in the near past it was unusual to hypothesize interfaces between acoustic models and pragmatic modules when describing the implementation of computational models of language analysis and generation. On the contrary, we claim that the complex interplay between the acoustic level and the pragmatic level of language constitutes an important aspect of voice expression of emotions. We also claim that the phonetic-pragmatics interplay can be accounted for in a text-to-speech system by providing accurate representations of contextually relevant discourse markers. The availability of an inventory of expressive cues implementing discourse markers, can improve the naturalness and expressivity of generated speech, moving toward the ambitious goal of emotional speech generation.

REFERENCES

[1] Akman V., Bazzanella C. (eds.) 2003 *Context*, special issue 35, *Journal of Pragmatics*, 321-504.
 [2] Bazzanella C. 2004 Emotions, Language and Context., In Weigand E. (ed.) 2004 *Emotion in dialogic interaction. Advances in the complex*. Amsterdam/Philadelphia, Benjamins., 59-76.
 [3] Bazzanella, C. 2006 Discourse Markers in Italian: towards a ‘compositional’ meaning. In Fischer K. (ed.) 2006. *Approaches to discourse particles*, Amsterdam, Elsevier, 504-524.
 [4] Brown P., Levinson S. 1978, 1987: *Universals in language usage: Politeness phenomena*, in E. N. Goody (ed.): *Questions and Politeness. Strategies in Social Interaction*. Cambridge, Cambridge University Press., 56-248; ed. 1987: *Politeness*. Cambridge UP, Cambridge.
 [5] Campbell, N. (2002), Towards a grammar of spoken language: incorporating paralinguistic information. In: 7th ISCA International

Conference on Spoken Language Processing, Denver, Colorado, USA, September 16-20, 2002.
 [6] Cresti, E. (2000), *Corpus di Italiano Parlato*. Volume I: Introduzione. Firenze, Accademia della Crusca.
 [7] Cresti, E. (2003), L’intonation des illocutions naturelles représentatives: analyse et validation perceptive, (<http://lablita.dit.unifi.it/publications/>)
 [8] Cresti, E. (2005), Per una nuova classificazione dell’illocuzione a partire da un corpus di parlato (LABLITA). In: Burr E. (Ed.), Atti del VI Convegno internazionale SILFI (giugno 2000, Duisburg), Cesati, Pisa.
 [9] Danieli, M., Emotional speech and emotional experience. *VIII International Conference of Neuro-Psychoanalysis (N-PSA): Neuro-Psychoanalytical Perspectives on Depression*, July 18-22, 2007, Wien.
 [10] A. Iida Akemi, N. Campbell, F. Higuchi & M. Yasumura, A corpus-based speech synthesis system with emotion. In: *Speech Communication*, Vol. 40, 2003: 161-187.
 [11] Hamza W., Bakis, R., Eide, E.M., Picheny, M. A., & Pitrelli, J. F. (2004), The IBM Expressive Speech Synthesis System. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Jeju, South Korea, October, 2004.
 [12] Johnstone, T. & Scherer, K.R., (1999). The effects of emotions on voice quality. In *Proceedings of XIV Int. Congress on Phonetic Science*.
 [13] Montero, J.M., Gutiérrez-Arriola, J.M., Palazuelos, S., Enríquez, E., Aguilera, S. & Pardo, J.M., Emotional Speech Synthesis: from Speech Database to TTS, In *Proceedings of ICSLP 1998*, Sidney.
 [14] Moore, E.I.I. Clements, M. Peifer, J *et al.*. Comparing objective feature statistics of speech for classifying clinical depression. In *Engineering in Medicine and Biology Society*, p17-20.
 [15] Murray, J.R. & Arnott, J.L., Synthesising emotions in speech: is it time to get excited?, In *Proceedings of ICSLP 96*, Philadelphia, pp.1816-1819.
 [16] Ochs E., Schieffelin B. 1989: *Language has a heart*, in E. Ochs (ed.): *The Pragmatics of Affect*, numero speciale di *Text* 9.1, 7-25.
 [17] Ochs E. Schegloff E.A., Thompson S.A. (eds.) 1996 *Grammar and Interaction*. Cambridge, Cambridge University Press.
 [18] Pally, R. (2001). A Primary Role for Nonverbal Communication in Psychoanalysis. In *Psychoanalytical Inquiry*, v21 p71-93.
 [19] Panksepp, J. (1999), Emotions as Viewed by Psychoanalysis and Neuroscience: An Exercise in Consilience, *Neuro-Psychoanalysis*, 1:15-38
 [20] Poggi, I. & Magno Caldognetto, E. (2004). Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici. In F. Albano Leoni, F. Cutugno, M. Pettorino & R. Savy. (Eds.), *Atti del Convegno “Italiano parlato”* (Napoli 14-15 febbraio 2003). Napoli: D’Auria Editore, CD-Rom.
 [21] Scherer, K.R. (2003), Vocal communication of emotion: A review of research paradigms, *Speech Communication*,
 [22] Schiffrin, D. 1987: *Discourse markers*. Cambridge, Cambridge University Press.
 [23] Schröder, M. (2001). Emotional Speech Synthesis: A Review, In *Proceedings of EUROSPEECH 2001*, pp. 561 – 564, Scandinavia, 2001.
 [24] Tini Brunozzi F., Quazza S. & Zovato, E., (to appear), Atti illocutivi e segnali discorsivi. Un contributo linguistico a un sistema TTS verso la sintesi vocale espressiva, atti del XL Congresso Internazionale di Studi della SLI “Linguistica e modelli tecnologici di ricerca”, Vercelli 21 - 23 settembre 2006, Roma, Bulzoni.
 [25] Zovato, E., Pacchiotti, A., Quazza, S. & Sandri, S., Towards emotional speech synthesis: a rule based approach. In: 5th ISCA Speech Synthesis Workshop, Pittsburgh USA, 2004.