Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication

Selma Yilmazyildiz, Wesley Mattheyses, Yorgos Patsis, and Werner Verhelst

Vrije Universiteit Brussel, dept. ETRO-DSSP
Pleinlaan 2, B-1050 Brussels, Belgium
{Selma.Yilmazyildiz, wmatthey}@vub.ac.be
{gpatsis, wverhels}@etro.vub.ac.be

Abstract. This paper presents our recent and current work on expressive speech synthesis and recognition as enabling technologies for affective robot-child interaction. We show that current expression recognition systems could be used to discriminate between several archetypical emotions, but also that the old adage "there's no data like more data" is more than ever valid in this field. A new speech synthesizer was developed that is capable of high quality concatenative synthesis. This system will be used in the robot to synthesize expressive nonsense speech by using prosody transplantation and a recorded database with expressive speech examples. With these enabling components lining up, we are getting ready to start experiments towards hopefully effective child-machine communication of affect and emotion.

1 Introduction

In Belgium alone some 300.000 children need to be hospitalized for long periods of time or suffer from chronic diseases [1]. Different projects exist which aim at using Information and Communication Technologies (ICT) like Internet and WebCams to allow these children to stay in contact with their parents, to virtually attend lectures at their school, etc. [1], [2]

Together with the Anty foundation and the Robotics and Multibody Mechanics research group at our university, we participate in a project that aims at designing a furry friendly robot called Anty [3], [4]. Anty will provide access to ICT means like a PC and WiMAX in a child-friendly form and will act as a friendly companion for the young hospitalized child. It is our task to design the vocal communication system for Anty. Since it will be a long time before a real speech dialog with a machine will become possible through speech understanding techniques, we choose to develop an affective communication system that can recognize expressive meaning in the child's voice, such as the child's intent or emotional state, and that can reply using synthesized affective nonsense speech.

The paper is organized as follows: in section 2 we describe our current emotion recognition system, in section 3 we describe our expressive synthesis system and in section 4 we conclude with a discussion.

2 Automatic classification of emotions in speech

It is well known that speech contains acoustic features that vary with the speaker's affective state. The effects of emotion in speech tend to alter pitch, timing, voice quality and articulation of the speech signal [5]. The goal of an emotional speech recognizer is to classify statistical measures of these acoustic features into classes that represent different affective states.

In our own work, we mainly used a segment based approach (SBA) for emotion classification. As illustrated in Fig. 1, statistical measures of acoustic features are calculated for the whole utterance as well as for each of its voiced segments. We used 12 statistical measures of pitch, intensity and spectral shape variation.

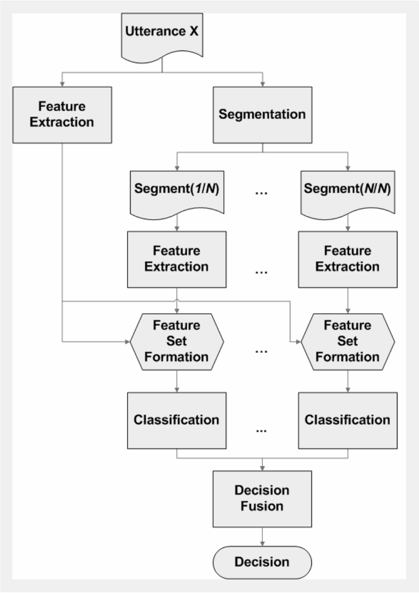

Fig. 1. Main components of the segment based approach for emotion classification.

Four different emotional databases (Kismet, BabyEars, Berlin and Danish) have been used and ten-fold cross validation has been mostly applied as the testing paradigm. Fig. 2 shows the results that we obtained with our SBA approach and with our own implementation of the AIBO emotion recognition system [6]. It can be noted that these results compare favourably to those that have been previously reported in the literature for these databases ([5], [7], [8], [9]).

In [10], we also reported some detailed cross database experiments. In these experiments the databases Kismet and BabyEars were paired and the emotional

	Kismet		BabyEars		Berlin		Danish	
MLA	AIBO	SBA	AIBO	SBA	AIBO	SBA	AIBO	SBA
SVM	83.7	83.2	65.8	67.9	75.5	65.5	63.5	56.8
KNN	82.2	86.6	61.5	68.7	67.7	59.0	49.7	55.6
ADA-C4.5	84.63	81	61.5	63.4	74.6	46.0	64.1	59.7

Fig. 2. Percentage recognition accuracy for emotion classification on four different databases with two different systems and three different machine learning algorithms.

classes that did not occur in both databases were dropped. The remaining common emotions were Approval, Attention and Prohibition. In a first set of experiments, training was performed on one of the databases and testing on the other. This off-corpus testing on the two corpora showed virtually no improvement over baseline classification (i.e., always classifying the test samples as belonging to the most frequent class in the test database). On the other hand, when the two corpora are merged into a single large corpus, classification accuracy is only slightly reduced compared to the scores obtained on the individual databases.

In other words, we found evidence suggesting that emotional corpora of the same emotion classes recorded under different conditions can be used to construct a single classifier capable of distinguishing the emotions in the merged corpora. The classifier learned using the merged corpora is more robust than a classifier learned on a single corpus because it can deal with emotions in speech that is recorded in more than one setting and from more speakers.

The emotional databases that are available contain only a small number of speakers and emotions and with the feature sets that are usually employed in the field, there is little generalization accross databases, resulting in database dependent classifiers. Furtunately, we also found that we can make the systems more robust by using much larger training databases. Moreover, adding robustness to the feature set that is used to represent the emotion in the utterance could compensate for the lack of vast amounts of training data. It would therefore be interesting to investigate the use of acoustic features that mimic the process of perception of emotions by humans.

3 Synthesis of affective nonsense speech

3.1 System design

We designed a system for producing affective speech that uses a database with natural expressive speech samples from a professional speaker and a database with naturally spoken neutral speech samples from the same speaker. Details of the construction of these databases are given in section 3.2.

In order to produce a nonsense utterance with a given desired emotion, the synthesizer randomly selects an expressive speech sample of the proper type from the first database and uses this as a prosodic template. Next, a nonsense carrier phrase is constructed that has the same syllabic structure as the selected prosodic template. As explained in detail in section 3.3, this is done by concatenating segments from the database with neutral speech samples. Except that inter-segment compatibility aspects are taken into account, the segments to be concatenated are selected ad random.

Finally, the same pitch and timing structure as found in the prosodic template is copied on the nonsense carrier phrase, a process that is known as prosodic transplantation [11], [12] and that effectively provides the synthetic output with a same intonational pattern as the natural example. The prosodic modification technique used in this prosody transplantation is summarized in section 3.4.

Besides working in accordance with the concept of prosodic transplantation, we believe that the strength of our synthesizer mainly resides in its high quality and low complexity that was achieved by using an overlap-add technique for both the segment concatenation and the prosodic modification, in accordance with the source filter interpretation of pitch synchronized overlap-add (PSOLA) [13], as introduced in [14]. As will be explained, according to this interpretation, the synthesizer can make use of the series of pitch markers to fulfill the concatenation. More details about the synthesizer can also be found in [15].

3.2 Speech database

In order to produce expressive speech as close as possible to natural emotional speech, the quality of the prosodic templates was an important parameter. The speaker should be able to keep a same voice quality while recording the neutral text and he should be able to express the desired emotions convincingly. Our databases were constructed from the speech samples of a professional speaker.

Four primary human emotions (anger, joy, sadness and fear) were included. First, samples of expressive and neutral utterances were recorded in an anechoic chamber. Next, the utterances to use in the databases were selected through an evaluation process using four criteria: color of the voice, the emotion perceived in the utterance, closeness to the intended emotion, and the quality of the portraying (faked/real). Each utterance was rated by four researchers who are familiar with speech processing and one amateur musician. Finally, 14 utterances were selected for inclusion in the database.

An interactive segmentation tool was developed based on the MEL-cepstral distances between the hanning windowed frames on the left and the right of each sample point. Large MEL-scale cepstral distances are an indication of a phone transition, small distances indicate stationary speech parts. The tool we developed plots these distances and, given the desired number of segments, the user can specify the appropriate cut-points. These are stored in a meta-data file that can be used for constructing the nonsense carrier utterances by concatenating randomly selected speech segments from the database. Fig. 3 illustrates this segmentation process.

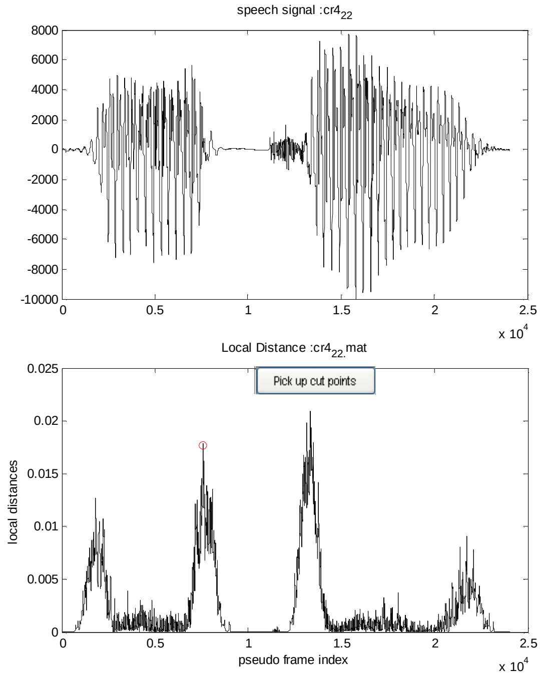

Fig. 3. MEL-cepstral distance based manual segmentation. The upper panel shows the speech utterance 'not ring'. The bottom panel shows the MEL-scale cepstral distances.

3.3 Segment concatenation

The synthesizer has to concatenate the selected segments in an appropriate way in order to construct a fluently sounding speech signal. While concatenating speech segments, one has to cope with two problems. First, the concatenation technique must smooth the transition between the two signals in time, otherwise these transitions will appear to abrupt and the concatenated speech would not sound fluent, but chopped. Further, while joining voiced speech signals, the introduction of irregular pitch periods at the transition point has to be avoided, since these would cause audible concatenation artifacts.

As mentioned before, we opted to use PSOLA to perform the prosody transplantation. PSOLA needs to identify the exact location of every individual pitch period in the voiced speech segments using so-called pitch markers. The quality of the output signal greatly depends on the correctness of these markers and we designed an efficient and robust algorithm to accomplish this pitch marking [16].

Obviously, by choosing pitch markers as the segments cut-points, we can assure that the periodicity of the speech signal will not be disrupted by the concatenation procedure. In order to further enhance the concatenation quality, we designed an optimization method that selects the best cut-markers according to a MEL-scale spectral distance, as suggested in [17]. This technique selects for each join a pitch marker from the first and from the second segment in such a way that the transition will occur where there is as much similarity between the two speech signals as possible.

Once the cut marks are determined, the actual concatenation problem is tackled by a pitch-synchronous window/overlap technique. First, a number of pitch periods (typically 5) is selected from the end cut-marker and from the beginning cut-marker of the first and second segment, respectively. Then, the pitch of these two short segments is altered using the PSOLA technique, which will result in two signals having exactly the same pitch. Finally, the two signals are cross-faded using a hanning-function to complete the concatenation.

Figure 4 illustrates our concatenation method by joining two voiced speech segments. To illustrate the method's robustness, we used a first segment that has a pitch value which is higher than that of the second segment, as one can see in the upper panel of the figure. The middle panel shows the pitch-alignment of the extracted pitch periods and the bottom panel shows the final concatenated speech. This last plot illustrates that in the concatenated speech signal the segment transition is smoothed among a few pitch periods, which is necessary if a fluent output is to be obtained. In addition, the output does not suffer from irregular pitch periods.

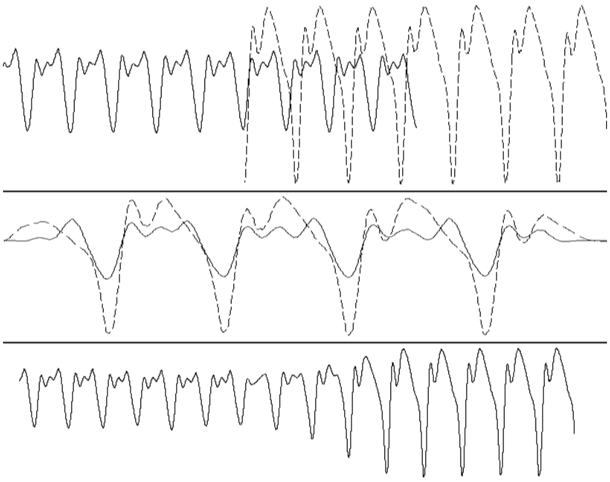

Fig. 4. Pitch-synchronous concatenation. The upper panel illustrates the segments to be concatenated, the middle panel illustrates the pitch-synchronized waveshapes, and the lower panel illustrates the result after cross-fading.

The proposed concatenation technique delivers results of the same auditive quality as some more complex concatenation methods found in the literature. The technique has been systematically judged against a spectral interpolation approach and it was concluded that the computationally more complex interpolation could not outperform the proposed overlap-add method.

3.4 Adding prosody

At this point we need to apply the correct prosody to the concatenated nonsense speech signal by using the PSOLA technique for altering the timing and the pitch of the speech. The pitch markers of the nonsense speech can be simply computed from the pitch markers of the concatenated segments. These will then be used as analysis-pitch markers for the PSOLA technique.

At the same time, each sample point that indicates a phoneme transition in the synthesizer's databases is memorized in the meta-data. By using these transition points the synthesizer calculates the inherent length of each phoneme present in the concatenated signal and in the prosodic template. Using these two sets of values, the amount of time-stretching that is necessary to provide the output speech with the correct timing properties is computed. Subsequently, the PSOLA algorithm will synthesize the output signal by using a time varying time-stretch value going from phoneme to phoneme. The synthesis-pitch markers used by the PSOLA operation determine the pitch of the final output [14]. Obviously, it suffices to calculate these pitch markers based on the pitch-parameters of the prosodic template to ensure that the imposed intonation curve is correctly assigned to the final speech signal.

4 Concluding discussion

We presented our recent and current work on expressive speech synthesis and recognition as enabling technologies for affective robot-child interaction.

We showed that our current expression recognition system obtains competitive results and could be used to discriminate between several archetypical emotions. However, we also showed that in this field the old adage "there's no data like more data" is more than ever valid and in order to avoid having to record hughe databases with expressive child speech, we plan to open a parallel track to investigate robust features for emotion recognition as well as psychoacoustically motivated dimensions of expressive speech.

We also designed a lightweight speech synthesis system that was successfully used as a replacement for the back-end of the NeXTeNS open source text-to-speech synthesis system for Dutch, thereby turning it into a Flemish speaking text-to-speech application [15]. We are using the same acoustic synthesis modules to construct a system for synthesizing expressive nonsense speech that copies the intonation from a database with expressive speech examples onto a neutral synthetic carrier phrase. In our future work we plan to investigate whether and how aspects of voice quality should be incorporated in the system.

With these enabling components lined up, we are getting ready to enter a new and very exciting research phase where we can start experiments towards hopefully effective child-machine communication of affect and emotion.

5 Acknowledgements

Parts of the research reported on in this paper were supported by the IWOIB project Link II - Voice Modification of the Brussels region, by the IWT projects SPACE (sbo/040102) and SMS4PA-II (O&O/040803), by the IBBT projects VIN and A4MC3, and by the research fund of the Vrije Universiteit Brussel.

References

1. Simon et Odil: Website for hospitalized children. http://www.simonodil.com/
2. IBBT research project ASCIT: Again at my School by fostering Communication through Interactive Technologies for long term sick children. https://projects.ibbt.be/ascit/
3. Anty project website: http://anty.vub.ac.be/
4. Anty foundation website: http://www.anty.org/
5. Breazeal, C., Aryananda, L.: Recognition of Affective Communicative Intent in Robot-Directed Speech. In: Autonomous Robots, vol. 12, (2002) pp. 83–104
6. Oudeyer, P.: The production and recognition of emotions in speech: features and algorithms. International Journal of Human-Computer Studies, Vol. 59 (2003) 157–183
7. Slaney, M., McRoberts, G. A Recognition System for Affective Vocalization. Speech Communication, 39 (2003) 367–384
8. Ververidis, D., Kotropolos, C.: Automatic speech classification to five emotional states based on gender information. Proceedings of Eusipco-2004 (2004) 341–344
9. Hammal, Z., Bozkurt, B., Couvreur, L., Unay, D., Caplier, A., Dutoit, T.: Passive versus active: vocal classification system. Proceedings of Eusipco-2005 (2005)
10. Shami, M., Verhelst, W.: Automatic Classification of Emotions in Speech Using Multi-Corpora Approaches. In: Proc. of the second annual IEEE BENELUX/DSP Valley Signal Processing Symposium SPS-DARTS (2006)
11. Verhelst, W., Borger, M.: Intra-Speaker Transplantation of Speech Characteristics. An Application of Waveform Vocoding Techniques and DTW. Proceedings of Eurospeech'91, Genova (1991) 1319–1322
12. Van Coile, B., Van Tichelen, L., Vorstermans, A., Staessen, M.: Protran: A Prosody Transplantation Tool for Text-To-Speech Applications. Proceedings of the International Conference on Spoken Language Processing ICSLP94, Yokohama (1994) 423–426
13. Moulines, E., Charpentier, F.: Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. Speech Communication, volume 9 (1990) 453-467
14. Verhelst, W.: On the Quality of Speech Produced by Impulse Driven Linear Systems. Proceedings of the International Conference on Acoustics, Speech and Signal Processing - ICASSP-91 (1991) 501–504
15. Mattheyses, W.: Vlaamstalige tekst-naar-spraak systemen met PSOLA (Flemish text-to-speech systems with PSOLA, in Dutch). Master thesis, Vrije Universiteit Brussel (2006)
16. Mattheyses, W., Verhelst, W., Verhoeve, P.: Robust Pitch Marking for Prosodic Modification of Speech Using TD-PSOLA. Proceedings of the IEEE Benelux/DSP Valley Signal Processing Symposium, SPS-DARTS (2006) 43–46
17. Conkie, A., Isard, I.: Optimal coupling of diphones. Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis - SSW2 (1994)