7 Conclusion
Automatic detection of emotions has been evaluated using spectral and pitch features, all modeled by GMMs on the frame level. Two corpora were used: telephone services and meetings. Results show that frame level GMMs are useful for emotion classification.

The two MFCC methods show similar performance, and MFCC-low outperforms pitch features. A reason may be that MFCC-low gives a more stable pitch measure. Also, it may be due to its ability to capture voice source characteristics, see Syrdal (1996), where the level difference between the first and the second harmonic is shown to distinguish between phonations, which in turn may vary across emotions.

The diverse results of the two corpora are not surprising considering their discrepancies. An possible way to improve performance for the VP corpus would be to perform emotion detection on the dialogue level rather than the utterance level, and also take the lexical content into account. This would mimic the behavior of the human labeler.

Above we have indicated the difficulty to compare emotion recognition results. However, it seems that our results are at least on par with those in Blouin & Maffiolo (2005).

Acknowledgements
This work was performed within CHIL, Computers in the Human Interaction Loop, an EU 6th Framework IP (506909). We thank Voice Provider for providing speech material.

References
3 Method and procedure
The prototype system was developed in several steps (see Figure 1). First, a Praat (Boersma & Weenink, 2005) script extracted 23 acoustic parameters every 10 ms. These were then used as input to the formant synthesiser GLOVE, which is an extension of OVE III (Liljencrants, 1968) with an expanded LF voice source model (Fant et al., 1985). GLOVE was used by kind permission of CTT, KTH. For a more detailed description, see Carlson et al. (1991).

Next, the parameters were adjusted to generate more natural-sounding synthesis. To be able to compare the natural speech to the synthetic versions, another Praat script was developed, which first called the parameter extraction script, and then displayed waveforms and spectrograms of the original word, the resulting synthetic word, as well as the previous synthetic version. By auditory and visual comparison of the three files, the user could easily determine whether a newly added parameter or adjustment had improved the synthesis. If an adjustment improved the synthesis, it was added to the adjustment rules. Formants, amplitudes and voice source parameters (except F0) caused the most serious problems, which were first solved using fixed values, then by parameter smoothing.

An attempt to synthesise speaker age was carried out using the system. The basic idea was to use the synthetic versions of the words to generate new words of other ages by age-weighted linear interpolation between two source parameter files. A Java program was developed to calculate the weights and to perform the interpolations. For each target age provided as input by the user, the program selects the parameter files of two source speakers (the older and younger speakers closest in age to the target age), and generates a new parameter file from the interpolations between the two source parameter files. For instance, for the target age of 51, i.e. exactly half-way between the ages of Speaker 2 (aged 36) and Speaker 3 (aged 66), the program selects these two speakers as source speakers, and then calculates the age weights to 0.5 for both source speakers. Next, the program calculates the target duration for each phone segment using the age weights and the source speaker durations. If the duration of a particular segment is 100 ms for Speaker 1, and 200 ms for Speaker 2, the target duration for the interpolation is 200 x 0.5 + 100 x 0.5 = 150 ms. All parameter values are then interpolated in the same way. Finally, the target parameter file is synthesised using GLOVE, and displayed (waveform and spectrogram) in Praat along with the two input synthetic words for comparison. A schematic overview of the procedure is shown in Figure 2.

4 Results
To evaluate the system’s performance, two perception tests were carried out to estimate direct age and naturalness (on a 7-point scale, where 1 is very unnatural and 7 is very natural). Stimuli in the first evaluation consisted of natural and synthetic versions of the 6, 36, 66 and 91 year old speakers. The second evaluation was carried out at a later stage when the 9, 39, 69 and 94 year olds had been included, and when parameter smoothing and pre-emphasis filtering (to avoid muffled quality) had improved the synthesis. 31 students participated in the first evaluation test, also including interpolations for 8 decades (10 to 80 years), while 21 students took part in the second, which also comprised interpolations for 7 decades (10 to 70 years).

4.1 First evaluation
In the first evaluation, the correlation curves between chronological age (CA, or simulated “CA” for the synthetic words) and perceived age (PA) displayed some similarity for the natural and synthetic words, though the synthetic ones were judged older in most cases, as seen in Figure 3. The interpolations were mostly judged as much older than both the natural and synthetic words. As for naturalness, the natural words were always judged more natural than the synthetic ones. Both the natural and synthetic 6 year old versions were judged least natural.

4.2 Second evaluation
Figure 4 shows that not only the correlation curves for the natural and synthetic words, but also for the interpolations did improve in similarity in the second evaluation compared to the first one. However, the natural and synthetic versions of the 39, 66 and 69 year olds were quite underestimated. All natural words were judged as more natural than the synthetic ones, and all synthetic words except the 6 and 94 year old achieved a median naturalness value of 6.
5 Discussion and future work

The synthetic words obtained a reasonable resemblance with the natural words in most cases, and the similarity in age was improved in the second evaluation. The interpolated versions were often judged as older than the intended age in the first evaluation, but in the second evaluation they had become more similar in age to the natural and synthetic versions, indicating that speaker age may be synthesised using data-driven formant synthesis. Still, some of the age estimations were quite unexpected. For instance, the 39, 66 and 69 year olds were judged as much younger than their CA. This may be explained by that these voices were atypical for their age.

One very important point in this study is that synthesis of age by linear interpolation is indeed a crude simplification of the human aging process, which is far from linear. Moreover, while some parameters may change considerably during a certain period of aging (i.e. F0 and formant frequencies during puberty), others remain constant. Better interpolation techniques will have to be tested. One should also bear in mind that the system is likely to interpolate not only between two ages, but also between a number of individual characteristics, even when the speakers are closely related.

Future work involves (1) improved parameter extraction for formants, (2) better interpolation algorithms, and (3) expansion of the system to handle more speakers (of both sexes), as well as a larger and more varied speech material. Further research with a larger material is needed to identify and rank the most important age-related parameters. If further developed, the prototype system may well be used in future studies for analysis, modelling and synthesis of speaker age and other speaker-specific qualities, including dialect and accent. The phonetic knowledge gained from such experiments may then be used in future speech synthesis applications to generate more natural-sounding synthetic speech.

References