it. Intensity peaks and F0 peaks correlate to some extent, yet it is clear that the two parameters should be treated separately. Note that speakers W and U have very similar F0 peak values but different intensity peak values.

4 Duration (speaking rate) and pause

Average total utterance duration for the three utterances for each speaker is presented in Figure 6. Of the three utterances, the utterance /a-soo-desuka/ permits the insertion of a pause after the initial interjection /a/. When pause duration is included, it shows the same durational pattern as the other two utterances without a pause in reflecting the attitude types. Therefore, we interpreted pause as part of durational manifestation and included it in total utterance duration. The smallest cross-speaker variation was found for NEU for which all except one speaker used the shortest duration, clustering around 600-800ms. In the absolute duration value, speakers were also uniform for SUS which falls in the range between 1000 to 1200ms. The greatest cross-speaker variation was found for DIS for which the duration of the utterance varied from 800ms to 1250ms.

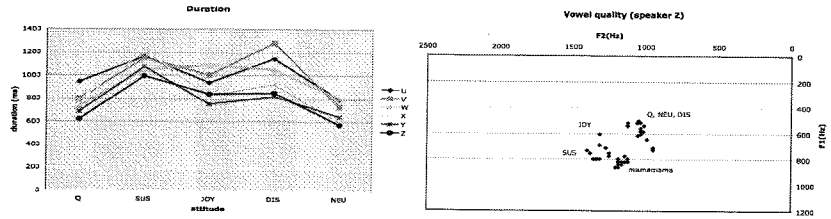

Figure 6 (left). Average utterance duration for each attitude type.
Figure 7 (right). Plotting of F1 and F2 for the vowel /a/ (speaker Z).

5 Vowel quality

Auditory impressions suggested considerable intra- and cross-speaker variation in the use of voice quality in general as well as in the specifically tested attitude types. Since the acoustic cues for voice quality are less straightforward than other acoustic cues, we only present the differences in vowel quality in this paper. Figure 7 above shows the manifestation of vowel quality by speaker Z. This speaker differentiated the vowel quality of /a/ in such a way that SUS and JOY had a more front quality than NEU, Q, and DIS. The figure also shows the formants values of /a/ in nonsense words /mamamama/ spoken neutrally by the same speaker.

6 Summary and discussion

Together with our earlier report on F0 shape and phrasing (Nagano-Madsen & Ayusawa 2005), both agreement and discrepancies were observable among the six speakers in their manifestation of attitudes. It seems that pragmatic information can be expressed in at least a few alternative ways in Japanese and that this line of research needs more attention.

References

Maekawa, K. & N. Kitagawa, 2002. How does speech transmit paralinguistic information? (in Japanese). *Cognitive Studies 9*(1), 46-66.

Nagano-Madsen, Y. & T. Ayusawa, 2005. Prosodic correlates of attitudinally-varied back channels in Japanese. *Proceedings FONETIK 2005*, Department of Linguistics, Göteborg University, 103-106.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 101–104

Emotion Recognition in Spontaneous Speech

Daniel Neiberg[1], Kjell Elenius[1], Inger Karlsson[1], and Kornel Laskowski[2]
[1]Department of Speech, Music and Hearing, KTH, Stockholm
{neiberg|kjell|inger}@speech.kth.se
[2] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA
kornel@cs.cmu.edu

Abstract

Automatic detection of emotions has been evaluated using standard Mel-frequency Cepstral Coefficients, MFCCs, and a variant, MFCC-low, that is calculated between 20 and 300 Hz in order to model pitch. Plain pitch features have been used as well. These acoustic features have all been modeled by Gaussian mixture models, GMMs, on the frame level. The method has been tested on two different corpora and languages; Swedish voice controlled telephone services and English meetings. The results indicate that using GMMs on the frame level is a feasible technique for emotion classification. The two MFCC methods have similar performance, and MFCC-low outperforms the pitch features. Combining the three classifiers significantly improves performance.

1 Introduction

Recognition of emotions in speech is a complex task that is furthermore complicated by the fact that there is no unambiguous answer to what the "correct" emotion is for a given speech sample (Scherer, 2003; Batliner et al., 2003). Emotion research can roughly be viewed as going from the analysis of acted speech (Dellaert et al., 1996) to more "real", e.g. from automated telephone services (Blouin & Maffiolo, 2005). The motivation of this latter is often to try to enhance the performance of such systems by identifying frustrated users.

A difficulty with spontaneous emotions is in their labeling, since the actual emotion of the speaker is almost impossible to know with certainty. Also, emotions occurring in spontaneous speech seem to be more difficult to recognize compared to acted speech (Batliner et al., 2003). In Oudeyer (2002), a set of 6 features selected from 200 is claimed to achieve good accuracy in a 2-person corpus of acted speech. This approach is adopted by several authors. They experiment with large numbers of features, usually at the utterance level, and then rank each feature in order to find a small golden set, optimal for the task at hand (Batliner et al., 1999).

Classification results reported on spontaneous data are sparse in the literature. In Blouin & Maffiolo (2005), the corpus consists of recordings of interactions between users and an automatic voice service. The performance is reported to flatten out when 10 out of 60 features are used in a linear discriminant analysis (LDA) cross-validation test. In Chul & Narayanan (2005), data from a commercial call centre was used. As is frequently the case, the results for various acoustic features were only slightly better than a system classifying all exemplars as neutral. Often authors use hundreds of features per utterance, meaning that most spectral properties are covered. Thus, to use spectral features, such as MFCCs, possibly with additional pitch measures, may be seen as an alternative. Delta MFCC measures on the utterance level have been used earlier, e.g. in Oudeyer (2002). However, we have chosen to model the distribution of the MFCC parameters on the frame level in order to obtain a more detailed description of the speech signal.

In spontaneous speech the occurrence of canonical emotions such as happiness and anger is typically low. The distribution of classes is highly unbalanced, making it difficult to measure and compare performance reported by different authors. The difference between knowing and not knowing the class distribution will significantly affect the results. Therefore we will include results from both types of classifiers.

2 Material

The first material used was recorded at 8 kHz at the Swedish company Voice Provider (VP), which runs more then 50 different voice-controlled telephone services. Most utterances are neutral (non-expressive), but some percent are frustrated, most often due to misrecognitions by the speech recognizer, Table 1. The utterances are labeled by an experienced, senior voice researcher into *neutral, emphasized* or *negative* (frustrated) speech. A subset of the material was labeled by 5 different persons and the pair-wise inter-labeler kappa was 0.75 – 0.80.

In addition to the VP data, we apply our approach to meeting recordings. The ISL Meeting Corpus consists of 18 meetings, with an average number of 5.1 participants per meeting and an average duration of 35 minutes. The audio is of 16 bit, 16 kHz quality, recorded with lapel microphones. It is accompanied by orthographic transcription and annotation of emotional valence (*negative, neutral, positive*) at the speaker contribution level (Laskowski & Burger, 2006). The emotion labels were constructed by majority voting (2 of 3) for each segment. Split decisions (one vote for each class) were removed. Finally, the development set was split into two subsets that were used for cross-wise training and testing.

Both corpora were split into a development and an evaluation set, as shown in Table 1.

Table 1. Materials used.

VP development set		
Neutral	3865	94 %
Emphatic	94	2 %
Negative	171	4 %
Total	4130	
VP evaluation set		
Neutral	3259	93 %
Emphatic	66	2 %
Negative	164	5 %
Total	3489	
ISL development set		
Neutral	6312	80 %
Negative	273	3 %
Positive	1229	16 %
Total	7813	
ISL evaluation set		
Neutral	3259	70 %
Negative	151	3 %
Positive	844	19 %
Total	4666	

3 Features

Thirteen *Standard MFCC* parameters were extracted from 24 Mel-scaled logarithmic filters from 300 to 3400 Hz. Then we applied RASTA-processing (Hermansky & Morgan, 1994). Delta and delta-delta features were added, resulting in a 39 dimensional vector. For the ISL material we used 26 filters from 300 to 8000 Hz; otherwise the processing was identical.

MFCC-low features were computed similarly to the standard MFCCs but the filters ranged from 20 to 300 Hz. We expected these MFCCs to model F0 variations.

Pitch was extracted using the Average Magnitude Difference Function, Ross et al. (1974) as reported by Langlais (1995). We used a logarithmic scale subtracting the utterance mean. Also delta features were added.

4 Classifiers

All acoustic features are modeled using Gaussian mixture models (GMMs) with diagonal covariance matrices measured over all frames of an utterance. First, using all the training data, a root GMM is trained with the Expectation Maximization (EM) algorithm with a maximum likelihood criterion, and then one GMM per class is adapted from the root model using the maximum a posteriori criterion (Gauvin & Lee, 1994). We use 512 Gaussians for MFCCs and 64 Gaussians for pitch features. These numbers were empirically optimized. This way of using GMMs has proved successful for speaker verification (Reynolds et al., 2000). The outputs from the three classifiers were combined using multiple linear regression, with the final class selected as the argmax over the per-class least square estimators. The transform matrix was estimated from the training data.

5 Experiments

We ran our experiments with the features and classifiers described above. An acoustic combination was composed by GMMs for MFCC, MFCC-low, and pitch. The combination matrix was estimated by first testing the respective GMM with its training data.

6 Results

Performance is measured as absolute accuracy, average recall (for all classes) and f1, computed from the average precision and recall for each classifier. The results are compared to two naïve classifiers: a random classifier that classifies everything with equal class priors, *random with equal priors*, and a random classifier knowing the true prior distribution over classes in the training data, *random using priors*. The combination matrix accounts for the prior distribution in the training data, heavily favoring the neutral class. Therefore a weight vector which forces the matrix to normalize to equal prior distribution was also used. Thus we report two more results: *acoustic combination* with equal priors, that is optimized for the accuracy measure and *acoustic combination using priors*, which optimizes the average recall rate. Thus, classifiers under the *random equal priors* heading do not know the a priori class distribution and should only be compared to each other. The same holds for the classifiers under *random using priors*. Note that the performance difference in percentages is higher for a classifier not knowing the prior distribution compared to its random classifier, than for the same classifier knowing the prior distribution compared to its random classifier. This is due to the skewed prior distributions.

Table 2. Results. Accuracy, Average Recall, f1.

VP Neutral vs. Emphasis vs. Negative			
Classifier	**Acc.**	**A.Rec.**	**f1**
Random with equal priors	*0.33*	*0.33*	*0.33*
MFCC	0.80	0.43	0.40
MFCC-low	0.78	0.39	0.37
Pitch	0.56	0.40	0.38
Acoustic combination	0.90	0.37	0.39
Random using priors	*0.88*	*0.33*	*0.33*
Acoustic comb. using priors	0.93	0.34	0.38
ISL Negative vs. Neutral vs. Positive			
Classifier	**Acc.**	**A.Rec.**	**f1**
Random with equal priors	*0.33*	*0.33*	*0.33*
MFCC	0.66	0.49	0.47
MFCC-low	0.66	0.46	0.44
Pitch	0.41	0.38	0.37
Acoustic combination	0.79	0.50	0.47
Random using priors	*0.67*	*0.33*	*0.33*
Acoustic comb. using priors	0.82	0.42	0.48

From Table 2 we note that all classifiers with equal priors perform substantially better than the random classifier. The MFCC-low classifier is almost as good as the standard MFCC and considerably better than the pitch classifier.

Regarding the ISL results in Table 2 we again notice that the pitch feature does not perform on the same level as the MFCC features. When the distribution of errors for the individual classes was examined, it revealed that most classifiers were good at recognizing the neutral and positive class, but not the negative one, most probably due to its low frequency resulting in poor training statistics.

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.

A 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

Batliner, A., J. Buckow, R. Huber, V. Warnke, E. Nöth & H. Niemann, 1999. Prosodic Feature Evaluation: Brute Force or Well Designed? *Proc. 14th ICPhS*, 2315-2318.

Batliner, A., K. Fischer, R. Huber, J. Spilkera & E. Nöth, 2003. How to find trouble in communication. *Speech Communication 40*, 117-143.

Blouin, C. & V. Maffiolo, 2005. A study on the automatic detection and characterization of emotion in a voice service context. *Proc. Interspeech*, Lisbon, 469-472.

Chul, M.L. & S. Narayanan, 2005. Toward Detecting Emotions in Spoken Dialogs. *IEEE, Transactions on Speech and Audio Processing 13*(2), 293-303.

Dellaert, F., T.S. Polzin & A. Waibel, 1996. Recognizing emotion in speech. *Proc. ICSLP*, Philadelphia, 3:1970-1973.

Gauvin, J-L. & C.H. Lee, 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. SAP 2*, 291-298.

Hermansky, H. & N. Morgan, 1994. RASTA processing of speech. *IEEE Trans. SAP 4*, 578-589.

Langlais, P., 1995. *Traitement de la prosodie en reconnaissance automatique de la parole.* PhD-thesis, University of Avignon.

Laskowski, K. & S. Burger, 2006. Annotation and Analysis of Emotionally Relevant Behavior in the ISL Meeting Corpus. *LREC*, Genoa.

Oudeyer, P., 2002. Novel Useful Features and Algorithms for the Recognition of Emotions in. Human Speech. *Proc. of the 1st Int. Conf. on Speech Prosody.*

Reynolds, D., T. Quatieri & R. Dunn, 2000. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing 10*, 19-41.

Ross, M., H. Shafer, A. Cohen, R. Freudberg & H. Manley, 1974. Average magnitude difference function pitch extraction. *IEEE Trans. ASSP-22*, 353-362.

Scherer, K.R., 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication 40*, 227-256.

Syrdal, A.K., 1996. Acoustic variability in spontaneous conversational speech of American English talkers. *Proc. ICSLP*, Philadelphia.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 105–108

Data-driven Formant Synthesis of Speaker Age

Susanne Schötz
Dept. of Linguistics and Phonetics, Centre for Languages and Literature, Lund University
susanne.schotz@ling.lu.se

Abstract

This paper briefly describes the development of a research tool for analysis of speaker age using data-driven formant synthesis. A prototype system was developed to automatically extract 23 acoustic parameters from the Swedish word 'själen' ['ɧɛːlən] (the soul) spoken by four differently aged female speakers of the same dialect and family, and to generate synthetic copies. Functions for parameter adjustment as well as audio-visual comparison of the natural and synthesised words using waveforms and spectrograms were added to improve the synthesised words. Age-weighted linear parameter interpolation was then used to synthesise a target age anywhere between the ages of 2 source speakers. After an initial evaluation, the system was further improved and extended. A second evaluation indicated that speaker age may be successfully synthesised using data-driven formant synthesis and weighted linear interpolation.

1 Introduction

In speech synthesis applications like spoken dialogue systems and voice prostheses, the need for voice variation in terms of age, emotion and other speaker-specific qualities is growing. To contribute to the research in this area, as part of a larger study aiming at identifying phonetic age cues, a system for analysis by synthesis of speaker age was developed using data-driven formant synthesis. This paper briefly describes the developing process and results.

Research has shown that acoustic cues to speaker age can be found in almost every phonetic dimension, i.e. in F_0, duration, intensity, resonance, and voice quality (Hollien, 1987; Jacques & Rastatter, 1990; Linville, 2001; Xue & Deliyski., 2001). However, the relative importance of the different cues has still not been fully explored. One reason for this may be the lack of an adequate analysis tool in which a large number of potential age parameters can be varied systematically and studied in detail.

Formant synthesis generates speech from a set of rules and acoustic parameters, and is considered both robust and flexible. Still, the more natural-sounding concatenation synthesis is generally preferred over formant synthesis (Narayanan & Alwan, 2004). Lately, formant synthesis has made a comeback in speech research, e.g. in data-driven and hybrid synthesis with improved naturalness (Carlson et al., 2002; Öhlin & Carlson, 2004).

2 Material

Four female non-smoking native Swedish speakers of the same family and dialect were selected to represent different ages, and recorded twice over a period of 3 years: *Speaker:1*: girl (aged 6 and 9), *Speaker 2*: mother (aged 36 and 39), *Speaker 3*: grandmother (aged 66 and 69), and *Speaker 4*: great grandmother (aged 91 and 94). The isolated word 'själen' ['ɧɛːlən] (the soul), was selected as a first test word, and the recordings were segmented into phonemes, resampled to 16 kHz, and normalized for intensity.