Classification of Affective Speech using Normalized Time-Frequency Cepstra

D. Neiberg[1], *P. Laukka*[2], *G. Ananthakrishnan*[1]

[1]Centre for Speech Technology (CTT), TMH, CSC, KTH, Stockholm, Sweden
[2]Department of Psychology, Stockholm University, Stockholm, Sweden
neiberg@speech.kth.se, petri.laukka@psychology.su.se, agopal@kth.se

Abstract

Subtle temporal and spectral differences between categorical realizations of para-linguistic phenomena (e.g., affective vocal expressions) are hard to capture and describe. In this paper we present a signal representation based on Time Varying Constant-Q Cepstral Coefficients (TVCQCC) derived for this purpose. A method which utilizes the special properties of the constant Q-transform for mean F0 estimation and normalization is described. The coefficients are invariant to segment length, and as a special case, a representation for prosody is considered. Speaker independent classification results using ν-SVM with the Berlin EMO-DB and two closed sets of basic (anger, disgust, fear, happiness, sadness, neutral) and social/interpersonal (affection, pride, shame) emotions recorded by forty professional actors from two English dialect areas are reported. The accuracy for the Berlin EMO-DB is 71.2 %, and the accuracies for the first set including basic emotions was 44.6% and for the second set including basic and social emotions the accuracy was 31.7% . It was found that F0 normalization boosts the performance and a combined feature set shows the best performance.
Index Terms: Emotion Classification, Constant-Q, 2D-DCT, supra-segmental, mean pitch estimation, prosody

1. Introduction

Human-computer interactions place lower cognitive loads on the human user when the interaction is more intuitive to the human. The detection of emotions in human speech by a dialog system, and corresponding adaptation of its response, may therefore improve the perception of naturalness and the intuitivity of the interface [1]. In this work, we focus on speaker independent classification in closed sets of basic and social/interpersonal emotions from acoustic measurements.

In recent years, the art of affective classification has been dominated by the method of doing machine learning from large sets of acoustic measurements. In the effort reported by [2], 4244 acoustic features from six laboratories were used to classify four affective states of German children interacting with a pet robot. These features were categorized into voice quality, fundamental frequency, spectral/formants, cepstral, wavelets, energy and duration. They reported that among these, wavelets gave a short-term multi-resolution of time, energy and frequencies in one and were thus superior in the modeling of temporal aspects.

The use of constant-Q MFCC and WPCC (Wavelet-Packet-Cepstral-Coefficients) was reported in a speaker independent cross-language emotion classification task [3]. The constant-Q means that the ratio of bandwidth and the center frequency of each band-pass filter in these filter-banks is constant. In a neutral to emotion classification task, they found that wavelets performed better than traditional MFCC, but no effort was made to compare standard equal bandwidth filters to the constant-Q filters.

The Berlin emotional database (EMO-DB) [4] is the most common corpus among authors reporting speaker independent results. It consist of 816 phrases, uttered by 10 professional actors, 5 female and 5 male. It consists of the emotions angry, sad, anxious, happy, bored and neutral. The distribution is not balanced, and the accuracy of recognizers can thus be enhanced by using prior distribution of classes. Often a subset is used, composed of all utterances ranked as at least 60% natural and at least 80% clearly percived in a preception test, which gives 494 utterances.

The temporal aspect was addressed by [5] on the EMO-DB. 276 features covering F0, energy, duration, formants, Harmonic-to-noise, MFCC, FFT, and zero crossing rate were used with SFFS and SVM. They further investigated three variants of sub-utterance-level features. Dividing phrases in three relative parts, still keeping the whole utterance statistics, gave remarkable performance improvement and the best results was 96.5% (it is unclear whether this result was speaker independent). Thus, temporal changes of features over a phrase are important.

The contribution of emotionally salient aspects of the F0 was investigated by [6] in a cross-cultural multi corpora task. First, pitch features in emotional speech were compared to neutral speech using a Kullback-Leibler distance. Second, the discriminative power of the features was derived using nestled regression models. They found that contour statistics on utterance level such as mean, maximum and minimum were more emotionally prominent than the curvature, which carried supplementary information at both sentence level and at smaller scale defined as segments of consecutive voiced frames.

In this work, we intend to derive a representation of the speech signal which is supra-segmental to cover temporal salience. Although wavelet based feature are attractive from a theoretical point of view, little effort has been made to make intuitive visualization based on these. Therefore, we start from theoretically optimal Constant-Q spectra and derive Time Varying Constant-Q Cepstral Coefficients as well as two desirable normalization techniques, one for the average fundamental frequency and another for utterance length. By using the fundamental frequency normalization, we construct a supra-segmental prosodic representation. The representation allows segments of varying length to be parametrized by a vector with a fixed size and therefore state-of-art classifiers such as Support Vector Machines may be used. Finally, an intuitive visualization technique for analysis based on *prototypical spectrograms* is demonstrated.

2. The Corpus

We utilized a selection of emotional utterances taken from the VENEC corpus [7]. Forty professional actors from USA and Australia vocally expressed various emotions (affection, anger, disgust, fear, happiness, pride, sadness, shame, and neutral) by enacting various emotion-eliciting situations. The actors were provided with scenarios describing typical situations in which each emotion may be elicited, based on current research on emotion appraisals. The verbal material consisted of one of two short phrases with emotionally neutral content (i.e., "Let me tell you something" or "That's exactly what happened") per expression. Our selection included a total of 360 emotional portrayals (40 speakers and 9 emotion categories, one sample per category). We also utilized the EMO-DB database as a reference [4].

3. Signal Processing

Spectro-temporal features are extracted from each utterance in the database first using a spectral transform to find the time-varying frequency components and then parameterized to find 2-dimensional Time Varying Constant-Q Cepstral Coefficients (TVCQCC).

3.1. The Spectral Transform

This paper uses a filter-bank based on the Constant-Q transform [8] with a corresponding Q factor of $1/(2^{1/12}-1)$ or 16.8 which corresponds to the 12 semitones per octave in a musical scale. The advantage of using such a filter-bank is convenient frequency normalization of an utterance as described in section 3.2. The k^{th} spectral component for the transform of the time signal $x(n) : 1 \leq n \leq N$ sampled at sampling frequency F_s are given by

$$X(k,n) = \sum_{m=1}^{L(k)} W_k(m)x(n-m)\exp\left(\frac{-j2\pi nC_f(k)}{F_s}\right) \quad (1)$$

where, $L(k)$ is the order/length of the window corresponding to the k^{th} spectral component. In a departure from the original paper [8], the windows function $W_k[m]$ are Finite Impulse Response (FIR) linear phase low pass filters. Their Central Frequencies (C_f) is given by $B * 2^{k/12}$, where B is the frequency (in Hertz) of the first spectral component and their Band-Widths (B_W) is given by

$$B_W(k) = B(2^{(k+1)/12} - 2^{(k-1)/12}) \quad (2)$$

and the order $L(k) = 2 * round(0.5 * F_s/B_W(k))$ The total number of filters in the bank are K, where $C_f(K)$ must be less than $F_s/2$. Thus the filter $W_k(n) : 1 \leq n \leq L(k)$ is given by

$$W_k(n) = \frac{\sin\left(\frac{(n-(L(k)/2))\,B_W(k)}{F_s}\right)}{\left(n-\frac{L(k)}{2}\right)} \quad (3)$$

Compared to Short-time Fourier Transform (STFT), the constant-Q transform has higher temporal resolution for higher frequencies and higher spectral resolution for lower frequencies, for the same number of filters. The varying bandwidths of the filters makes the frame shift rate less important for the Constant-Q transform compared to STFT, and here a frame shift rate of 250 Hz is used.

3.2. Mean Frequency Estimation and Normalization

The method proposed in this paper is not a true fundamental frequency detection algorithm, but an approximation of the same for the purpose of finding the mean pitch in an utterance. It is based on a popular frequency domain method following Klapuri [9]. The Constant-Q filter-bank outputs are frequencies grouped in bins corresponding to the semitones of the musical scale instead of a linear scale. The i^{th} harmonic, $N_h(i)$, relative to the fundamental frequency is found by the following equation [10]

$$N_h(i) = round(12\log_2 i) \quad (4)$$

This would be far more complicated to do in a perceptual scale like Mel-scale. The first 12 harmonics are considered because beyond that consecutive harmonics would fall under the same bin. An approximation for tone in noise separation is used here which classifies all frequencies with amplitudes below 10 dB from the highest amplitude frequency component as noise. So any local maximum above this threshold occurring in the output of the filter-bank is considered as tones, which means that the summing starts at the first index $\hat{k}$ containing non-noise. For every frequency bin, the frequency amplitudes of the harmonically related frequencies are summed for the entire utterance (F_a).

$$F_a(k,n) = \sum_{i=1}^{12} |X(k+N_h(i),n)|^2 \quad (5)$$

The estimate for the instantaneous pitch (F_0) is the maximum among the summed harmonic frequency bins.

$$F_0(n) = \arg\max_{1\leq k\leq K} \quad F_a(k,n) \quad (6)$$

The mean fundamental frequency M_{ff} (in semitone scale) of the utterance is calculated from this estimate but weighted by the amplitude of the harmonic frequency sum.

$$M_{ff} = \frac{1}{\sum_{n=1}^{N}(F_0(n),n)} \sum_{n=1}^{N} F_0(n)F_a(F_0(n),n) \quad (7)$$

Normalizing the sentence to an arbitrary mean fundamental frequency N_{ff} (in the semitone scale), which is possible over the entire database, is simply done by shifting the filter-bank outputs corresponding to the difference between M_{ff} and N_{ff}.

$$X_n(k,n) = X(k+M_{ff}-N_{ff},n) \quad (8)$$

where, X_n is the normalized output of the filter-bank.

3.3. The Time Varying Constant-Q Cepstral Coefficients

The most commonly used acoustic parameterizations used for speech recognition and recently in synthesis are the Mel Frequency Cepstral Coefficients (MFCC). Cepstra are often calculated by taking the Cosine transform of the short time logarithm spectrum of the acoustic signal. It is known that MFCC of consecutive segments of speech are highly correlated. In order to tap the time-varying information, velocity (or acceleration) coefficients are often added in the parameterizations. A two-dimensional cepstrum along the time and frequency dimensions has been suggested by Ariki *et. al.* [11] as an alternative for expressing time-variation. This was done using a linear frequency scale. It was later adapted to the Mel Frequency scale by Milner and Vaseghi [12]. Such a parameterization of speech is shown to be a time-varying representation with highly de-correlated coefficients.

In this paper, we propose a form of time varying parametrization over entire utterances. This has not been tried before in the emotion recognition tasks, to the authors knowledge. The proposed method provides a suitable way of integrating the information available in the entire sentence into a matrix of fixed size. The formulation is as follows. First the log time-varying spectrum is obtained

$$lX(k,n) = 10\log_{10}(|X(k,n)|^2) \tag{9}$$

Then, the TVCQCC are calculated by applying a 2 dimensional discrete cosine transform (2D-DCT), as follows. For $1 \leq p \leq P$ and $1 \leq q \leq Q$, (where P and Q are the number of coefficients in the frequency and time dimension respectively), the TVCQCC are

$$T(p,q) = \sum_{n=1}^{N}\sum_{k=1}^{K} \frac{lX(k,n)}{N} * cos\left(\frac{\pi(k-\frac{1}{2})(p-1)}{K}\right) * cos\left(\frac{\pi(n-\frac{1}{2})(q-1)}{N}\right) \tag{10}$$

The axis of T along q is called the 'quefrency' and has a time dimension. The axis along p is the frequency of quefrency, which here is referred to as 'meti' (following the convention of swapping syllables), and has frequency dimension. It should be noted that the 2D-DCT has been modified so that this representation is length invariant, which means that the parameters are not affected by stretching or compression in time. In that sense, this representation is normalized to utterance length.

4. Experiments and Results

The experiments are designed to validate the mean F0 estimate against a reference, evaluate the benefit of mean frequency normalization and to assess the contribution of F0 alone compared to the full frequency range. Three sets of affective classes are chosen as:

1. Emo_A: anger, disgust, fear, happiness, sadness (Basic emotions) and neutral
2. Emo_B: anger, disgust, fear, happiness, sadness, affection, pride, shame (Basic and social/interpersonal emotions) and neutral
3. Emo_{DB}: angry, sad, anxious, happy, bored and neutral (from the Berlin corpus)

As default, we use a filter-bank starting at $B = 60$ Hz and ending at 5753.5 Hz, with 12 filters per octave, spanning a total of 79 bins, denoted as FB_{Full}. For the prosody representation, the first 35 bins (60-453 Hz) from FB_{FULL} are kept, creating FB_{F0}. The correlation between the mean frequency M_{ff} and the mean frequency obtained by Praat [13] (both variables are recalculated to Hertz) is 0.88 ($p << 0.01$) for set Emo_B. The default frequency for normalization is chosen as $N_{ff} = 23$ (226.5 Hz), which is 1 octave below the maximum frequency of FB_{F0}. Thus, variations in F0 up to one octave above the mean F0 are kept. The normalized full frequency filter bank is denoted FB_{NFull}, while the normalized F0-filter bank, denoted FB_{NF0}, is created by keeping the first 35 bins from FB_{NFull} (rather than from FB_{Full}). While using 12 quefrency cepstra coefficients is a common practice, it is reasonable to overshoot and let the machine learner extract the necessary patterns, but for the 'meti' coefficients there is no common practice and we chose a reasonable number based on inspection of the prototypical spectrograms. Thus, for for filter-banks FB_{Full} and FB_{NFull}, we use $Q = 30$ quefrency and $P = 30$ meti coefficients, and for filter-banks FB_{F0} and FB_{NF0}, we use $Q = 20$ quefrency and $P = 30$ meti coefficients. In addition, the cepstra vectors of FB_{NFull} and FB_{NF0} are merged, creating FB_{Comb}.

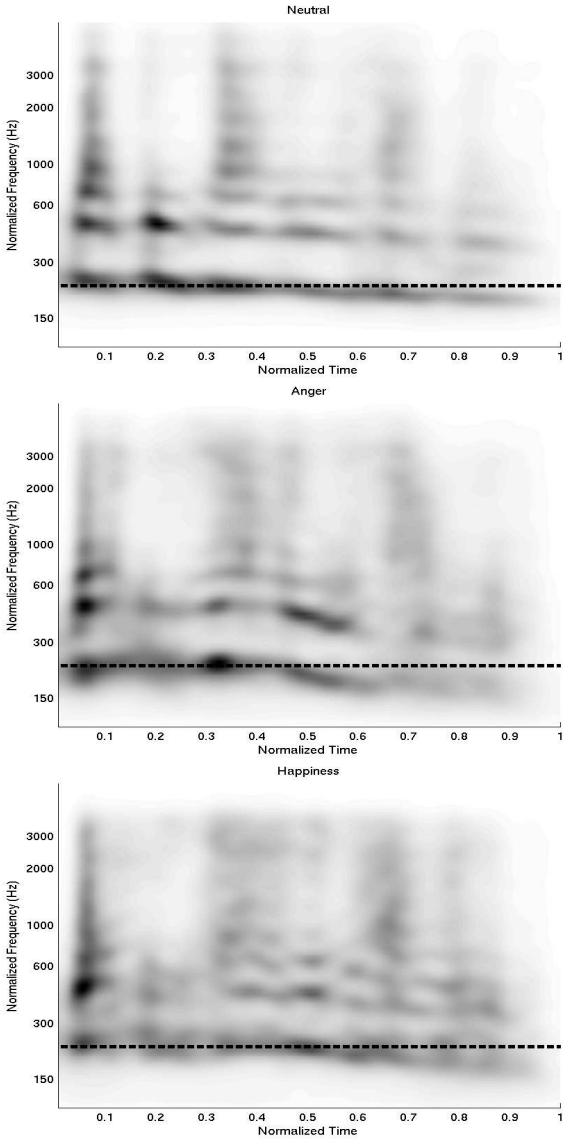

Figure 1: *Prototypical spectrograms* of neutral at the top, anger in the middle and happiness at the bottom. The dashed line shows the normalized mean frequency. Compared to neutral, the F0 of anger has larger but slowly moving variation and happiness has faster moving variation.

Figure 1 shows *prototypical spectrograms* with power amplitude scale obtained from calculating the average TVCQCC of all instances of three classes in Emo_A using FB_{NFull} followed by inverse transformation.

The TVCQCC matrices are converted into vectors by stacking the rows after each other, then the utterance length is added to the vector. If F0 normalization is applied, then the estimated mean pitch is added. Classification of the resulting vectors is done by using ν-SVM with linear kernel from the LibSVM package [14], with default shrinking heuristics. We choose to do a leave-one-out evaluation scheme on speaker level, and the

ν parameter is optimized using cross-validation on each training set. The results are shown in Table 1 and 2. We report accuracy, average recall for the unbalanced Emo_{DB} and F1-score which is the harmonic mean of precision and recall. Standard deviation is given in parentheses.

Table 1: *Classification results for Emo_A and Emo_B. Accuracy and F1-score is given in percentages.*

Set	Emo_A	Emo_A	Emo_B	Emo_B
Measure	Acc.	F1	Acc.	F1
FB_{FULL}	39.2 (3.2)	39.1 (3.1)	29.4 (2.4)	28.9 (2.4)
FB_{NFULL}	44.2 (3.2)	44.0 (3.2)	30.6 (2.4)	30.0 (2.4)
FB_{F0}	31.2 (3.0)	29.7 (3.0)	25.6 (2.3)	24.9 (2.3)
FB_{NF0}	40.0 (3.2)	39.4 (3.2)	28.1 (2.4)	27.0 (2.3)
FB_{Comb}	44.6 (3.2)	44.5 (3.2)	31.7 (2.5)	31.1 (2.4)

Table 2: *Classification results for Emo_{DB}. Accuracy, F1-score and average recall is given in percentages.*

Measure	Acc.	F1	avg R.
FB_{FULL}	66.6 (2.0)	64.0 (2.1)	63.8 (2.1)
FB_{NFULL}	69.7 (2.0)	66.9 (2.0)	66.3 (2.0)
FB_{F0}	60.7 (2.1)	58.4 (2.1)	58.1 (2.1)
FB_{NF0}	67.5 (2.0)	65.2 (2.1)	64.2 (2.1)
FB_{Comb}	71.2 (2.0)	68.8 (2.0)	68.1 (2.0)

5. Discussion

Both FB_{Full} and FB_{F0} benefited from normalization, and the F0 representation was almost as efficient as the full frequency representation when F0-normalization was applied. Further, the combined feature set showed the best performance. For the Emo_{DB}, Wagner et. al., [15] reported accuracy of 73.92% and an average recall of 61.36%, Li et.al., reported accuracy of 69.1% [16], and Lugger et. al. [17] reported 66.7% accuracy which they increased to 74.5% in a two-step approach. These results are comparable to our accuracy of 71.2 % and average recall of 68.1%, despite the fact that we did not use feature selection. For Emo_A and Emo_B, it is hard to make a fair comparison of the results to other works, but the relative low accuracy may be explained by the absence of perceptive post-selection of stimuli, and the more non-prototypical nature of the stimuli. It should be noted, however, that the accuracy was approximately three times higher than chance also for these sets, which is comparable to the average results usually obtained with human listeners.

6. Conclusions

In this work, we have outlined a supra-segmental signal representation referred to as Time Varying Constant-Q Cepstral Coefficients (TVCQCC). The coefficients were invariant to segment length, and a simple fundamental frequency normalization procedure which utilized the special properties of the constant-Q transform was described. As a special case, a representation for prosody was shown. Three different closed sets of basic and social/interpersonal emotions were used for speaker independent classification with ν-SVM. It was shown that the special F0 representation was almost as efficient as the full frequency representation, and the combined features showed the best performance. We hope to use this novel approach to model also other paralinguistic phenomena, and are further investigating techniques for analysis based on *prototypical spectrograms* as shown in Figure 1 .

7. Acknowledgments

This work was partly founded by the Swedish Research Council under contract 2006-1360.

8. References

[1] Huber, R., Batliner, A., Buckow, J., Nth, E., Warnke, V., and Niemann, H., "Recognition of emotion in a realistic dialogue scenario," in Proc. Int. Conf. on Spoken Language Processing, 665–668, 2000.

[2] Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V., "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in Interspeech 2007. Interspeech, 2007.

[3] Kandali, A. B., Routray, A., and Basu, T. K., "Vocal emotion recognition in five native languages of assam using new wavelet features," International Journal of Speech Technology, 12(1):1–13, March 2009.

[4] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B., "A database of german emotional speech," in Interspeech 2005, Lissabon, 1517–1520, 2005.

[5] Schuller, B. and Rigoll, G., "Timing levels in segment-based speech emotion recognition," in INTERSPEECH 2006, Pittsburgh, USA, September 2006.

[6] Busso, C., Lee, S., and Narayanan, S., "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," Audio, Speech, and Language Processing, IEEE Transactions on, 17(4):582–596, May 2009.

[7] Laukka, P., Elfenbein, H. A., Chui, W., Thingujam, N. S., Iraki, F. K., Rockstuhl, T., and Althoff, J., "Presenting the venec corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals," in LREC 2010 Workshop on Corpora for Research on Emotion and Affect, 2010.

[8] Brown, J., "Calculation of a constant Q spectral transform," J Acoust Soc of Am, 89(1):425–434, 1991.

[9] Klapuri, A., "Multiple fundamental frequency estimation by summing harmonic amplitudes," in Proc. ISMIR, 216–221, 2006.

[10] Ananthakrishnan, G., "Music and speech analysis using the 'bach' scale filter-bank," M.S. thesis, Indian Institute of Science, 2007.

[11] Ariki, Y., Mizuta, S., Nagata, M., and Sakai, T., "Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum," Communications, Speech and Vision, IEE Proceedings I, 136(2):133–140, Apr 1989.

[12] Milner, B. and Vaseghi, S., "An analysis of cepstral-time matrices for noise and channel robust speech recognition," in Fourth European Conference on Speech Communication and Technology. ISCA, 1995.

[13] Boersma, P. and Weenink, D., *Praat: doing phonetics by computer*, 2008, Software available at `http://www.praat.org/`.

[14] Chang, C.-C. and Lin, C.-J., *LIBSVM: a library for support vector machines*, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[15] Wagner, J., Vogt, T., and André, E., *Affective Computing and Intelligent Interaction*, chapter A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech, 114–125, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, September 2007.

[16] Fu, L., Mao, X., and Chen, L., "Speaker independent emotion recognition using hmms fusion system with relative features," in Intelligent Networks and Intelligent Systems, 2008. ICINIS '08. First International Conference on, 608–611, November 2008.

[17] Lugger, M. and Yang, B., "The relevance of voice quality features in speaker independent emotion recognition," in Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, 4:IV–17–IV–20, April 2007.