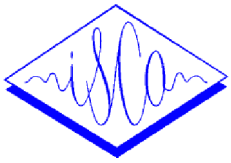
ISCA Archive
http://www.isca-speech.org/archive



Odyssey 2008: The Speaker and
Language Recognition Workshop
Stellenbosch, South Africa
January 21-24, 2008

Age and Gender Classification using Modulation Cepstrum

Jitendra Ajmera and Felix Burkhardt***

Deutsche Telekom Laboratories*
T-Systems Enterprise Services GmbH**
Berlin, Germany
{jitendra.ajmera, felix.burkhardt}@telekom.de

Abstract

This paper proposes using modulation cepstrum coefficients instead of cepstral coefficients for extracting meta-data information such as age and gender. These coefficients are extracted by applying discrete cosine transform to a time-sequence of cepstral coefficients. Lower order coefficients of this transformation represent smooth cepstral trajectories over time. Results presented in this paper show that cepstral trajectories corresponding to lower (3-14 Hz) modulation frequencies provide best discrimination. The proposed system achieves 50.2% overall accuracy for this 7-class task while accuracy of human labelers on a subset of evaluation material used in this work is 54.7%.

1. Introduction

The task of determining age and gender of a speaker given a short speech utterance is a challenging task and has gained significant attention recently. A system to detect elderly voice based on prosodic features *jitter* and *shimmer* was presented in [1]. Cepstral coefficients with acoustic modeling techniques of speakers were applied for the task of differentiating *subjectively* old speakers from others in [2]. AT&T's *how may I help you* (HMIHY) system detects age and gender among various other *voice signatures* [3]. An insight into phonetic knowledge about correlates to speaker age was provided in [4] where *classification and regression trees* (CART) was used for predicting the age of a speaker. A comparison of four approaches for this task was presented in [5].

Speech based age and gender analysis has various important applications. This classification can be useful in targeted commercial promotions for products and services companies. Analysis of user groups for market research can be facilitated with such classification. Age and gender information can be used to adapt degree of automation, manner and order of presentation of a dialogue in an *interactive voice response* (IVR) system. This analysis can also serve as an important tool toward making adaptive and novel interfaces for elderly people [6, 7]. Speech recognition technology can also benefit as acoustic models can be selected or adapted on the fly. Alternatively, an appropriate vocal tract length normalization transformation on the standard feature vectors can be applied for the purpose of speaker independent speech recognition.

Cepstral coefficients [8] have been investigated by many researchers for the purpose of age and gender discrimination [2, 3, 5]. This is potentially inspired by the fact that these coefficients work well for a variety of other speech related problems such as speech and speaker recognition. It can be argued that while for most speech problems the actual content of the signal is important and characterized by static cepstral coefficients, meta-data information such as age/gender should be best characterized by a slowly moving temporal envelope of the signal.

With this motivation, in this work we investigate using smooth representation of cepstral trajectories over time for the task of age and gender classification. In summary, discrete cosine transform (DCT) is applied to time-sequence of cepstral coefficients over a context window and first few DCT coefficients are retained to represent smooth cepstral trajectories. This can also be viewed as filtering in modulation cepstrum domain. Results presented in this paper show that modulation frequencies as low as 3-14 Hz are enough for extracting age and gender information.

This paper is organized as follows: Section 2 explains mathematical notations and derivation of smooth cepstral trajectories. Section 3 explain the database, experiments and evaluation used to show effectiveness of cepstral trajectories compared to mel frequency cepstral coefficients (MFCC) features [8].

2. Modulation Cepstrum

Different representations and derivations of cepstral trajectories have been presented in the literature [9, 10, 11] for the purpose of speech recognition. The basic idea is to decompose temporal sequences of cepstral coefficients into a set of modulation frequency components from where slow and fast varying factors can be selected.

If $C(n,k)$ denote the k^{th} cepstral coefficient at frame n, smooth cepstral trajectories can be viewed as output of Q bandpass filters (centered around relatively lower mod-

ulation frequencies) applied to the sequence $C(n,k), n = 1...P$. These features are referred to as mel cepstral modulation spectrum (MCMS) features in [9] and are characterized as cepstral time matrices in [10, 11].

MCMS coefficients are extracted by applying discrete cosine transform (DCT) to $C(n,k), n = 1...P$. Accordingly, the q^{th} MCMS coefficient $MCMS(n,k,q)$ of k^{th} cepstral coefficient at time n is computed as:

$$MCMS(n,k,q) = \sum_{p=0}^{P-1} C(n,k) cos \frac{\pi q(p+0.5)}{P} \quad (1)$$

with $q \in \{0, P-1\}$ and $k \in \{0, K-1\}$, where K is the number of cepstral coefficients extracted at every time frame n, and P is the length of the context window used to extract MCMS features.

As mentioned earlier, this DCT operation can also be viewed as filtering in the modulation cepstrum domain and lower coefficients ($q \in \{1, Q\}, Q < P-1$) represent slowly varying cepstral trajectories. Concatenating Q such coefficients for every cepstral coefficient would result into a vector of dimension $Q \cdot K$ at every frame n.

A comprehensive derivation and explanation of these features was provided in [9]. It was observed that cepstral modulation frequencies in the range 3-22 Hz were useful for speech recognition and higher modulation frequencies generally deteriorated the performance. Our results in Section 3.3 show that "meta-data" information like age and gender can be extracted using even lower (3-14Hz) modulation frequencies.

Also, smooth **cepstral** trajectories are preferred compared to smooth **spectral** trajectories [12] for the reason that spectral energies in adjacent bands are highly correlated [9].

The next section presents an evaluation framework that was used to show and compare performance of MCMS features for age and gender classification.

3. Experiments and Evaluations

The experimental framework, database and evaluation technique used in this work is similar to the one used in [5]. Since the training and test-sets used are exactly the same, the results presented in this paper are directly comparable to the ones presented in [5].

The task consists of classifying a speech utterance into the following 7 groups and labels:

- Children: ≤ 13 years (C)
- Young people: 14-19 years, male (YM) and female (YF)
- Adults: 20-64 years, male (AM) and female (AF)
- Seniors: ≥ 65 years, male (SM) and female (SF)

3.1. Database

The evaluation data was taken from the German SpeechDat II corpus [13], which is annotated with age and gender labels as given by callers at the time of recording. This database consists of 4000 native German speakers who called a recording system over the telephone and read a set of numbers, words and sentences. 80 speakers of each age and gender group were selected for training and 20 for testing, thereby gaining a weighted age and gender structure. The training data consisted of the whole utterance set of each person, up to 44 utterances.

In order to evaluate the performance on data that originates from a different domain, the system was also evaluated on the VoiceClass data. This data consists of 660 native speakers of German, which called a voice recorder and freely talked for about 5 to 30 seconds on the topic of their favorite dish. The age structure is not controlled, the data consists of many children and youth but almost no seniors.

Finally, the performance was evaluated on a total of more than 6000 utterances, approximately balanced among 7 classes. The complexity of the task can be realized by the fact that overall classification accuracy of human labelers on a subset of evaluation material used in this work is only 54.7% [5].

The human labeling experiments are based on 30 listeners classifying randomly selected 100 test utterances from SpeechDat corpus with no context information. Overall classification accuracy is 54.7%. The difference between long and short sentences also exists for human labelers, although human labelers do not perform that much worse on short sentences (60% and 50.9%).

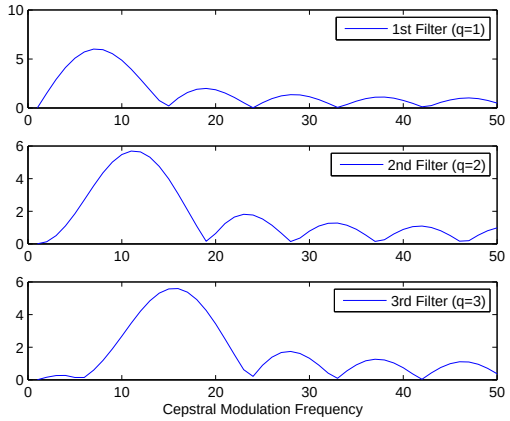
3.2. Pre-Processing

A simple energy based speech detector was first applied to extract speech segments and $K = 12$ MFCC features were extracted every 10ms from these segments. Cepstral mean substraction (CMS) was applied at utterance basis to remove channel effects.

For the proposed system, $Q = 3$ MCMS coefficients were extracted (Eq. 1) over a context window of $P = 11$ frames (120ms), resulting into 36 coefficients. For comparison, delta and double-delta coefficients were used with MFCC features, also resulting into 36 coefficients.

In this work, $C(n,k)$ are sampled at 100 Hz (every 10ms) and P MCMS coefficients characterize modulation frequencies from 0 to 50 Hz. Therefore, first 3 MCMS coefficients represent modulation frequencies from 3-14 Hz approximately.

256 component Gaussian mixture model (GMM) was trained for each of the 7 classes for both MFCC and MCMS features after K-Means initialization. For testing, log-likelihood scores were computed for each class given a speech utterance and class generating maximum



ss

Figure 1: Modulation frequency response of the first 3 MCMS filters. As seen here, these filters emphasize different modulation frequencies providing complementary information.

likelihood was assigned to that utterance.

3.3. Evaluation and Comparison

Tables 1 and 2 present classification accuracies obtained when using MFCC (plus delta and double delta) and MCMS features, respectively. In these tables of results, horizontal rows show the percentages of utterances belonging to a particular class and sum up to 100. It is clear that MCMS features provide much better performance (50.2% overall accuracy) compared to MFCC features (45.7% overall accuracy). This suggests that for extracting meta-data information like age and gender, a slowly varying temporal envelope of speech signal is more useful than the static information that is generally used for speech recognition.

Another explanation for better performance of MCMS features is provided by the observation that while delta and double delta coefficients emphasize the same cepstral modulation frequency around 15 Hz, MCMS filters emphasize different modulation frequencies providing complementary information. This is illustrated in Figure 1.

Figure 2 shows classification accuracy with different number of DCT coefficients (Q) used, which is equivalent to allowing different ranges of modulation cepstrum frequencies. It can be concluded that $Q = 3$ DCT coefficients provide maximum classification accuracy and higher order coefficients tend to deteriorate the performance. This is potentially due to the fact that higher order coefficients (fast varying components) provide more details about the actual *content* of the signal which is not really necessary for age and gender information.

Further analysis was carried out to investigate the impact of utterance length on the classification accuracy. A sub-set of short utterances SpeechDat short (SpeechDat II corpus identifiers 'a' and 'o', average duration 3.5 seconds) and another subset of longer sentences SpeechDat long (identifier 's', average duration 8 seconds) were evaluated for this purpose. This analysis is presented in Table 3. As expected, classification accuracy for short utterances is much lower compared to those of long utterances.

Class	C	YM	YF	AM	AF	SM	SF
C	55	1	23	6	13	1	1
YM	0	45	1	32	4	15	3
YF	17	0	47	0	27	0	9
AM	0	14	0	51	2	33	0
AF	4	2	13	1	43	5	32
SM	1	19	0	30	2	42	6
SF	6	3	5	1	41	4	40

Table 1: Confusion matrix (rows sum up to 100) and classification accuracies of different classes using MFCC features. Overall accuracy is 45.7%.

Class	C	YM	YF	AM	AF	SM	SF
C	65	2	17	4	6	2	4
YM	1	55	1	21	4	17	1
YF	28	0	44	0	20	0	8
AM	0	21	0	48	1	30	0
AF	9	1	13	1	45	3	28
SM	1	13	1	23	1	54	7
SF	7	1	6	1	35	3	47

Table 2: Confusion matrix (rows sum up to 100) and classification accuracies of different classes using MCMS features. Overall accuracy is 50.2%.

The overall accuracy obtained using MCMS features is superior to the performance obtained using other features/methods on exactly the same task [5] except that of an approach which utilizes a phoneme recognizer. This can be expected since a phoneme recognizer utilizes many more sources of information and results into a much more complex system.

4. Conclusions

Modulation cepstrum coefficients which represent smooth cepstral trajectories are used in this work in

Overall	SD_short	SD_long	VoiceClass
50.2%	41.4%	56%	57%

Table 3: Classification accuracy for different subsets of test-data based on utterance lengths.

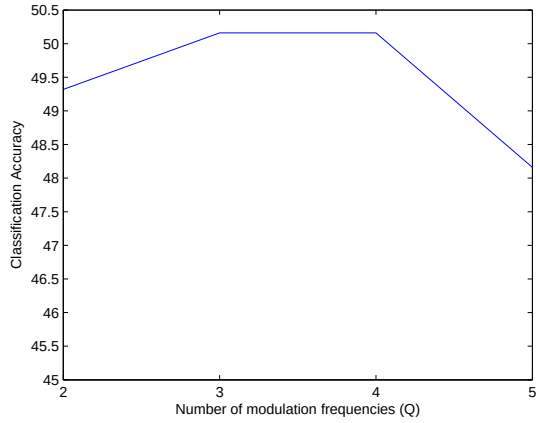


Figure 2: Classification accuracy with varying number of modulation frequency components (Q). It can be seen that modulation frequencies as low as 3-14 Hz (first 3 components) are enough for age and gender discrimination.

place of actual cepstral coefficients for the task of age and gender classification. Discrete cosine transform is applied to a sequence of cepstral coefficients and the first few DCT coefficients are retained for each cepstral dimension. This can also be seen as filtering operation in cepstral modulation domain. It is shown that these features provide much better classification accuracy (50.2%) for a 7-class task compared to cepstral coefficients (45.7%). Comparison of different number of modulation frequency coefficients showed that modulation frequencies as low as 3-14 Hz are enough for extracting age and gender information.

5. References

[1] Christian Mueller, Frank Wittig and Joerg Baus, "Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs", Eurospeech, 2003.

[2] Minematsu N., Sekiguchi M, and Hirose K., "Automatic estimation of ones age with his/her speech based upon acoustic modeling techniques of speakers", ICASSP, 137–140, 2002.

[3] I. Shafran, M. Riley and M. Mohri, " Voice Signatures ", Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2003.

[4] Susanne Schoetz, "Automatic prediction of speaker age using CART", Term paper for course in Forensic Phonetics, Goeteborg University.

[5] Florian Metze et. al. "Comparison of Four Approaches to Age And Gender Recognition for Telephone Applications", ICASSP 2007.

[6] J. A. Jorge, "Adaptive tools for the elderly: new devices to cope with age-induced cognitive disabilities,", EC/NSF workshop on Universal accessibility of ubiquitous computing, ACM Press, 2001, pp. 66-70.

[7] C. Mueller and R. Wasinger,, "Adapting Multimodal Dialog for the Elderly", ABIS-Workshop 2002 on Personalization for the Mobile World, 2002.

[8] Davis S., and Mermelstein P., "Comparison of parametric representations for monosyllablic word recognition in continuously spoken sentences ", IEEE Trans. Acouc., Speech, Signal Proc. (ASSP), 28:357–366, 1980.

[9] V. Tyagi, I. McCowan, H. Misra and H. Bourlard, "Mel-Cepstrum Modulation Spectrum (MCMS) Features For Robust ASR", ASRU 2003.

[10] B.P. Milner and S.V. Vaseghi, "An analysis of cepstral time feature matrices for noise and channel robust speech recognition", Proc. Eurospeech, pp. 519-522, 1995.

[11] B.P. Milner, "Inclusion of temporal information into features for speech recognition", Proc. ICSLP, PP. 256-259, 1996.

[12] B.E.D. Kingsbury, N.Morgan and S. Greenberg, "Robust speech recognition using the modulation spectrogram", Speech Communication, vol. 25, Nos. 1-3, August 1998.

[13] European Language Resources Association (ELRA), http://www.speechdat.org/, http://www.elra.info/.