

Time-Varying Cepstral Coefficients

Trond Skogstad, Torbjørn Svendsen

Department of Electronics and Telecommunications,
Norwegian University of Science and Technology

tronds@iet.ntnu.no, torbjorn@iet.ntnu.no

Abstract

This paper introduces a new set of cepstral features, based on a slightly modified version of the time-varying linear predictive models pioneered by Subba Rao and Liporace. In these models, the non-stationarity of the speech signal is accommodated by expressing the filter coefficients as a weighted combination of known basis functions. By running the parameterized filter coefficients through the recursive link between all-pole models and cepstral coefficients, we obtain a time-varying cepstral representation in analytical form. In a preliminary recognition experiment this representation is shown to give a satisfactory performance. It is argued that the introduced features are well suited for tasks such as detection of landmarks and stationary segments.

Index Terms: time-varying linear prediction models, stationarity, ASR front-end

1. Introduction

All prevailing front-ends for ASR treat the speech signal as a sequence of fixed-length stationary segments. It is not. Rather, the phones comprising the speech signal have a highly variable duration and is at best quasi-stationary for some part of the duration. Some phones are never acoustically stable, meaning information must be found in the transients.

Recognizing this, there has been an increased interest in landmark detection [1] and the related problem of detecting stationary segments [2]. These are two step approaches to the estimation of feature vectors, the first step supplying the second step information on e.g. segment boundaries and the nature of the segment under consideration. This way the reliability of the features could be increased by tailoring the choice of features and allowing the use of the maximum possible number of samples in estimating these features. In this first step of feature extraction, the dynamics of the speech signal is especially important.

It is in this setting we introduce a new description of the speech signal, time-varying cepstral coefficients. As the regular cepstral coefficients, they provide a representation of the log magnitude spectrum of the signal, but are here presented in an analytical form with the same temporal resolution as the sampled signal itself. This means a description of the dynamics inside an analysis frame could be made, whereas conventional cepstral parameters only allow for an analysis of the dynamics between frames. In this respect, they appear well suited for use in the first step of a two step feature estimation procedure. It is also possible to use the representation directly as features for speech recognition allowing longer frames than what is appropriate with traditional features such as MFCC and LPCC.

These time-varying cepstral coefficients are derived from time-varying linear predictive models, where the usual assumption

of stationarity is replaced with the assumption that the time-development of the filter coefficients is expressible using a known set of basis functions. This is of course a limiting premise, but it allows for the estimation of the model parameters using linear algebra.

2. Time-varying linear prediction

Linear predictive (LP) analysis has been one of the mainstays of speech technology for four decades due to its concise representation of the speech signal and the relative ease of extracting the necessary parameters. In the classical formulation of the problem [3], the signal x_t is assumed to be stationary over an analysis frame and approximated as a linear combination of the past P samples

$$\hat{x}_t = a_1 x_{t-1} + a_2 x_{t-2} + \dots + a_P x_{t-P}. \quad (1)$$

The coefficients are found by minimizing the l_2 -norm of the difference $\{e_t\}$ between $\{x_t\}$ and $\{\hat{x}_t\}$. Several authors have tried to extend this model by letting the filter coefficients have a temporal development inside the analysis frame, thereby allowing some degree of non-stationarity. Seemingly, the most popular approach is to express the coefficients as a weighted combination of known basis functions, that is:

$$a_k(t) \doteq \sum_{i=0}^M a_{ik} u_i(t), \quad k = 1, \dots, P. \quad (2)$$

The implicit assumption is no longer of stationarity, rather that the statistical properties of the signal are well behaved (changes sufficiently slowly) and are expressible in the chosen basis. In the first published effort on the subject [4], Subba Rao used the second-order Taylor expansion of the coefficients. Liporace [5] also used powers of time, of arbitrary order, and provided efficient ways of computing the filter coefficients. Other polynomial bases that have been tried are the Legendre functions and the discrete prolate spheroidal sequences [6]. Hall et al. [9] experimented with both polynomial bases and trigonometric functions (Fourier series), noting only minor differences in the resulting time trajectories of the filters. In this work the first few functions from the DCT-II are used. This provides a slowly varying basis, suited for estimation, and will have advantages for the cepstral description of the signal presented in Section 3.

Other efforts in the field of time-varying linear prediction include [7], where the authors combine LPC techniques with singular value decomposition and obtain a representation where the prediction filter is allowed to move linearly with time inside a frame, and [8], where joined estimation of adjacent frames are used to find a representation where the coefficients evolve linearly within the frame and are connected continuously between

frames. Using methods similar to the ones discussed in Section 3, a cepstral representation could be derived for both sets of parameters.

2.1. The time-varying all-pole model of speech

By allowing the coefficients a_k to have a time-development as specified by eq. (2), we get the following model for the signal x_t

$$\hat{x}_t = a_1(t)x_{t-1} + a_2(t)x_{t-2} + \dots + a_P(t)x_{t-P} + e_t. \quad (3)$$

Summarizing the derivation of the optimal (in the least mean square sense) coefficients from [5], we define

$$x \doteq \{x_t\}, \hat{x}_t \doteq \{\hat{x}_t\}, s_{ik} \doteq \{u_i(t)x_{t-k}\}, e \doteq \{e_t\}, \quad (4)$$

all being vectors of length equal to the length of the analysis frame, T . Also,

$$r_{ij}(k, l) = \langle s_{ik}, s_{jl} \rangle, \quad i, j = 0, \dots, M; \quad k, l = 1, \dots, P. \quad (5)$$

Further, R_{ij} denotes the matrix whose (k, l) element is $r_{ij}(k, l)$ and $r_{0j} \doteq (r_{0j}(0, 1), \dots, r_{0j}(0, P))'$. The filter coefficients are grouped such that $A = (A_0, \dots, A_M)' = (a_{01}, \dots, a_{0P}, \dots, a_{M1}, \dots, a_{MP})'$. The norm square of the error vector is

$$\|e\|^2 = \|x\|^2 - 2 \langle x, \hat{x} \rangle + \|\hat{x}\|^2 \quad (6)$$

$$= \|x\|^2 - 2 \sum_{i=0}^M A_i' r_{0i} + \sum_{i=0}^M \sum_{j=0}^M A_i' R_{ij} A_j \quad (7)$$

$$= \|x\|^2 - 2A'r_0 + A'RA \quad (8)$$

$$= \|x\|^2 - r_0'R^{-1}r_0 + (RA - r_0)'R^{-1}(RA - r_0) \quad (9)$$

where R is an $(M+1)P \times (M+1)P$ symmetric matrix with R_{ij} as its (i, j) -th sub-block and $r_0 = (r_{00}, \dots, r_{0M})'$. Since $A'RA$ is the norm square of $\|\hat{x}\|$ and thus strictly non-negative, the matrix R and thereby R^{-1} is positive definite, and the minimum of eq. (9) is found when $RA - r_0 = 0$, that is when

$$A = R^{-1}r_0. \quad (10)$$

2.2. Inclusion of a gain term

In a certain way, the method presented in section 2.1 presupposes a constant gain in the model or equally a signal with roughly constant power. To illustrate this, consider an analysis frame spanning two distinct segments of speech, one high energy vowel and one low energy fricative. If the theoretically attainable prediction gain in the two segments are of similar proportions, the minimization of the squared sum of errors will cause the use of almost all degrees of freedom in the description of the vowel segment, to the detriment of the description of the fricative segment - that is, a small improvement of the prediction in the high energy segment leads to a bigger error reduction than a large improvement of the prediction in the low energy segment.

To counteract this effect, we propose to modify the error term in eq. (3) from e_t to $g_t e_t'$, where the gain g_t will represent the systematic variations in power in the signal and e_t' represents the prediction error given the gain. If the gain is assumed known, a suitable object for minimization is the norm square of the modified error vector e' . By defining W as a diagonal

matrix with $\frac{1}{g_t}$ as its elements, this can be seen to be equal to a weighted minimization of the norm square of the original error term, e_t :

$$\|e'\|^2 = \|We\|^2 = \|e\|_W^2. \quad (11)$$

If we redefine $s_{ik} \doteq \{\frac{1}{g_t} u_i(t)x_{t-k}\}$ and all correlation terms $r_{ij}(k, l)$ are calculated using this new definition, the problem is equivalent to the one in section 2.1:

$$\|e'\|^2 = \|Wx\|^2 - 2 \langle Wx, W\hat{x} \rangle + \|W\hat{x}\|^2 \quad (12)$$

$$= \|Wx\|^2 - 2 \sum_{i=0}^M A_i' r_{0i} + \sum_{i=0}^M \sum_{j=0}^M A_i' R_{ij} A_j, \quad (13)$$

and the same procedure for finding the coefficients apply.

While this method has evident benefits, it also has some unfavorable implications. Consider the stationary case. It is well known [3] that the criteria of minimal $\|e\|^2$ equals the criteria of a maximally flat residual spectrum $E(\omega)$. When minimizing $\|e'\|^2$, this is no longer the case. The multiplication of e_t in the time domain with a weight function w_t corresponds to the convolution $E(\omega) * W(\omega)$ in the frequency domain. We thus have two irreconcilable wishes for the function w_t : 1. It is responsible for weighting the error samples according to the gain of the system, and as such should be non-flat in time. 2. It should be an impulse in frequency, to make sure the residual spectrum is maximally flat. This is obviously impossible, but it is still feasible to find weight functions that do a satisfactory job in the time domain, while keeping most of its energy inside a narrow main lobe in the frequency domain.

In an effort to keep the mathematics tractable, the determination of g_t (and thereby the weight function) and e_t are done in separate steps. To find g_t , the analysis frame of length T is subdivided in smaller frames of length T_{sub} . In each such frame, a stationary all-pole model is fitted to the signal and the gain is calculated. To smooth the gains, making sure the weight function has a low-pass frequency response, the gains from each frame is projected to a cosine basis and the first few terms are kept. The gains are then interpolated to T samples by changing the time basis of the cosine functions. The weights are determined as $w_t = \frac{1}{g_t}$ - the implication being that the predictive powers of the stationary (constant) all-pole models are similar to the predictive powers of the time-varying model in each sub-frame. If this is true the modified error signal e_t' , calculated using the time-varying model, should have variance close to unity in all subframes.

2.3. Stability of models

A serious drawback with the time-varying model, probably responsible for the limited use of the model in the literature and real systems, is the difficulty in guaranteeing stability. When the model parameters are found by solving eq. (10), there is a chance the filters will be unstable at some of the time instances. In [10], a new method for estimating the parameters that guarantees stability is presented. However, the method departs quite drastically from the original statement as a linear estimation problem and is considerably more complex. If the models are to be used in ASR, the stability problem could be bypassed with heuristics such as detecting unstable frames and reestimating after adding small amounts of white noise, or by instructing the back-end recognizer to disregard frames that are detected as unstable.

3. Cepstral features calculated from time-varying LP models

The cepstral coefficients have been shown to be more reliable and robust features for ASR than the LP filter coefficients and other parameters derived from the LP filters, such as the PARCOR coefficients and the log area ratio coefficients. In this section it will be shown that it is possible to calculate cepstral features directly from time-varying LP models. A new feature set for speech recognition will also be introduced: The weights of the cosine basis functions that expresses the cepstral coefficients.

If the all-pole model

$$H(z) = \frac{G}{1 - \sum_{k=1}^P a_k z^{-k}} \quad (14)$$

has all poles inside the unit circle (is stable), the real cepstrum can be computed directly from the LP filter coefficients by the recursion [11]:

$$\hat{h}[n] = \begin{cases} 0 & n < 0 \\ \ln G & n = 0 \\ a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) \hat{h}[k] a_{n-k} & 0 < n \leq p \\ \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) \hat{h}[k] a_{n-k} & n > p \end{cases} \quad (15)$$

Now, focusing on $0 < n \leq p$ and assuming the filter coefficients are expressed using cosine basis functions, that is,

$$a_k(t) = \sum_{i=0}^M a_{ik} u_i(t) = \sum_{i=0}^M a_{ik} d_i \cos(\omega_i t), \quad k = 1, \dots, P \quad (16)$$

we postulate that the n -th cepstral coefficient at time t will be expressible in the form

$$\hat{h}[n, t] = \sum_{l=0}^{L(n)} \beta_{nl} \cos(\Omega_{nl} t). \quad (17)$$

Applying (16) and (17) to (15), we get

$$\begin{aligned} \hat{h}[n, t] &= \sum_{i=0}^M a_{in} d_i \cos \omega_i t \\ &+ \sum_{k=1}^{n-1} \binom{k}{n} \sum_{l=0}^{L(k)} \beta_{kl} \cos(\Omega_{kl} t) \sum_{i=0}^M a_{n-k,i} d_i \cos(\omega_i t) \end{aligned} \quad (18)$$

$$\begin{aligned} &= \sum_{i=0}^M a_{in} d_i \cos \omega_i t \\ &+ \sum_{k=1}^{n-1} \sum_{l=0}^{L(k)} \sum_{i=0}^M \binom{k}{n} \beta_{kl} a_{n-k,i} d_i \cos(\Omega_{kl} t) \cos(\omega_i t) \end{aligned} \quad (19)$$

$$\begin{aligned} &= \sum_{i=0}^M a_{in} d_i \cos \omega_i t \\ &+ \sum_{k=1}^{n-1} \sum_{l=0}^{L(k)} \sum_{i=0}^M \binom{k}{2n} \beta_{kl} a_{n-k,i} d_i \phi(t), \end{aligned} \quad (20)$$

where

$$\phi(t) = \cos(\Omega_{kl} - \omega_i)t + \cos(\Omega_{kl} + \omega_i)t. \quad (21)$$

Since at no step in the recursion non-cosine terms are introduced, it's evident that the postulated form in eq. (17) is correct.

By grouping the terms according to their frequency and using the applicable cosine symmetry properties, e.g $\cos(-\theta) = \cos(\theta)$, the coefficients β_{nl} and the number of terms $L(n)$ can be calculated and stored for use in later steps of the recursion.

The expression for calculating $\hat{h}[n, t]$ when $n > p$ is the same as for $0 < n \leq p$, with the exception of the a_n -term, and it can thus be calculated using the same procedure, skipping the first sum. The 0-th cepstral coefficient $\hat{h}[0, t]$ has to be treated separately, since the solution of eq. (3) in itself does not produce a error sequence in an analytical form. One possible solution is to use the gain terms found for weighting purposes in section 2.2.

An interesting characteristic of the proposed cepstral coefficients is that the number of basis functions needed to express the coefficients are increasing with increasing order (increasing n). The frequencies Ω_{nl} of these basis functions are also increasing. The cepstral coefficients are thus able to change more rapidly than the filters coefficients from which they are derived. If the DCT-II basis is used, the number of basis functions necessary for expressing the n -th cepstral coefficient is $n(M+1)$. However, the variance of β_{nl} is rapidly decreasing for increasing l , making the error introduced by truncation negligible if e.g. only the $M+1$ first coefficients are kept. As for conventional cepstral coefficients, the variance of $\hat{h}[n, t]$ is also decreasing for increasing n . The cepstral representation introduced in this paper is thus able to represent the spectrum with a reasonable number of parameters.

3.1. Properties of the cepstral features

Now let's examine some interesting properties of the proposed cepstral features of eq. (17). First of all, if no truncation is done, the cepstral features $\hat{h}[n, t]$ of eq. (17) evaluated at time t are exactly equal to the ones found by evaluating the filter coefficients in eq. (16) at time t and using the standard recursion in eq. (15). If the signal model in eq. (3) is accepted as a true model of the speech signal, that is, the filters found correspond to some true configuration of the speech production apparatus at all time instances t , the proposed cepstral parameters retain all properties of conventional cepstral parameters, presented with the same time resolution as the signal itself.

Since the time derivative of the log magnitude spectrum of the signal x_t has the Fourier series representation [11]

$$\frac{\partial}{\partial t} [\log |X(e^{j\omega}, t)|] = \sum_{m=-\infty}^{\infty} \frac{\partial h[m, t]}{\partial t} e^{-j\omega m} \quad (22)$$

and our proposed cepstral coefficients are expressed in an analytical form, with trivially obtainable time derivatives, finding the spectral rate of change is easy. This can be used to answer questions such as "how stationary is the signal at time t ?" and "at which time does the maximum spectral change occur?", and as such can be a valuable tool in segmentation- and landmark detection-tasks. Similarly, other questions relating to the movement of energy in the time-frequency plane could be answered.

4. Experiments

As a proof of concept, we have performed some preliminary recognition experiments. The goal was to demonstrate that the proposed features contain a similar amount of information on the nature of the speech signal analyzed as conventional MFCCs. The chosen task was recognition on the “A-set” (consisting of 30 isolated application words) of the Norwegian SpecDat database. As a baseline the COST 249 SpeechDat reference recognizer [13] was used, using monophone (M) and triphone (T) models with 8, 16 and 32 Gaussian mixture components and a front-end configuration with the 13 first MFCCs and their Δ s and $\Delta\Delta$ s. The window length used was 25ms and the frame shift 10 ms.

The same back-end recognizer was used for the recognition experiments with the proposed features (PF). The frame length was 100 ms and the frame shift 20 ms. The frame length was chosen to span approximately the same range as the MFCCs (when including the Δ s and $\Delta\Delta$ s), though the analysis is obviously less focused on the central part of the frame. The estimation of the time-varying models was done with $M=3$ and $P = 12$. The feature vectors presented to the back-end recognizer consisted of the 36 parameters β_{nl} with $n = \{1, \dots, 12\}$ and $l = \{0, 1, 2\}$ and 3 parameters representing the gain, calculated as the third root of the first three coefficients in the cosine projection of the sub-frame gains from section 2.2. As both systems use 39 features, the comparison is fair in terms of model parameters. Approximately 0.5% of the frames were found to be unstable during the cepstral recursion. In this experiment, these frames were simply discarded.

Since the proposed features are intended for applications where the time-evolution of the speech signal is most important, the same experiment was repeated using only the 26 dynamic features (MFCC DF and PF DF) for both systems.

Table 1: Word error rate for the recognition experiment on the A-set of SpeechDat

	MFCC	PF	MFCC DF	PF DF
M8GM	13.03	13.54	22.57	23.25
M16GM	11.07	10.99	20.27	18.82
M32GM	9.37	9.80	17.80	15.76
T8GM	4.68	5.11	6.30	5.88
T16GM	4.51	4.68	4.43	5.45
T32GM	4.94	5.45	4.51	4.86

In most of the cases the proposed features performs slightly worse than the MFCC baseline. The proposed features performs relatively better for monophone models than for triphone models. An informal comparison between the confusion matrices showed some differences in the errors made by the MFCC-systems and the systems using the proposed features. It should be noted that no great effort was put in optimizing the free parameters of the proposed system. It could well be the cepstral coefficients should be truncated in another way, e.g. using relatively more features to represent the lower order cepstral coefficients, which are known to be less spurious than the higher order cepstral coefficients. Different frame lengths and frame shifts could also prove to be better. As we are comparing our new system with a highly optimized system, the results are quite encouraging.

5. Conclusions

We have proposed a new set of cepstral features based on time-varying linear predictive models. To accommodate frames with varying power, the estimation procedure was modified to include a time dependent gain term. In a recognition experiment, the features were shown to give results comparable to conventional MFCC parameters. Since the representation is not based on an assumption of stationarity, the time-varying cepstral coefficients provide an interesting characterization of the dynamics of the speech signal inside an analysis frame. Further, the features have the same temporal resolution as the signal itself and are presented in an analytical form, making them well suited for use in detection of landmarks and stationary segments.

6. References

- [1] Espy-Wilson, C.Y., Pruthi, T., Juneja, A., Deshmukh, O., “Landmark-based Approach to Speech Recognition: An alternative to HMMs”, Proc. of Interspeech 2007, Antwerpen, pp. 886-889
- [2] Tyagi, V., Boulard, H., Wellekens, C., “On variable-scale Piecewise Stationary Spectral Analysis of Speech Signals for ASR”, Speech Communication, Vol. 48, Issue 9, September 2006, pp. 1182-1191.
- [3] Markel, J.D, Gray Jr., A.H., Linear Prediction of Speech, Springer-Verlag, Berlin, 1976.
- [4] Subba Rao, T., “The fitting of nonstationary time-series models with time-dependant parameters”, J. Royal Statist. Soc. Series B, Vol 32, No.2, 1970,pp.312-322.
- [5] Liporace, L.A., “Linear estimation of nonstationary signals”, J. Acoust. Soc. Am., Vol 58, No.2, 1975, pp.1288-1295.
- [6] Grenier, Y., “Time-Dependent ARMA Modeling of Nonstationary Signals”, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-31, No.4, Aug. 1983
- [7] L.F.M. ten Bosch, “Time-Varying Singular Value Decomposition: An application on speech-like signals”, Proc. of ICASSP 1991, Toronto, pp. 3197-3200, vol. 5
- [8] Schnell, K., Lacroix, A., “Time-Varying Linear Prediction for Speech Analysis”, Proc. of EURASIP 2007, Poznan, pp. 2045-2049.
- [9] Hall, M., Oppenheim, A.V., Willsky, A., “Time-Varying Parametric Modeling of Speech”, Proc.IEEE Decision Contr. Conf, New Orleans, 1977, pp. 1085-1091
- [10] Juntunen, M., Tervo, J., Kaipio, J.P., “Stabilization of Subba Rao-Liporace Models”, Circuits Systems Signal Process., Vol. 18., No. 4, 1999, pp. 395-406
- [11] Huang, X., Acero, A., Hon, H-W., “Spoken Language Processing”, Prentice Hall PTR, New Jersey, 2001.
- [12] Rabiner, L., Juang B.H, “Fundamentals of Speech Recognition”, Prentice-Hall PTR, New Jersey, 1993
- [13] Johansen, F.T, Warakagoda, N., Lindberg, B., Lehtinen, G., Kacic, Z., Zgank, A., Elenius, K., Salvi, G., “The COST 249 SpeechDat Multilingual Reference Recognizer”, Second International Conference on Language Resources and Evaluation (LREC-2000),May,2000