ROBUST SPEECH FEATURE EXTRACTION BY GROWTH TRANSFORMATION IN REPRODUCING KERNEL HILBERT SPACE

Shantanu Chakrabartty, Yunbin Deng and Gert Cauwenberghs

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218, USA
{*shantanu,yunbin,gert*}*@jhu.edu*

ABSTRACT

A robust speech feature extraction procedure, by kernel regression nonlinear predictive coding, is presented. Features maximally insensitive to additive noise are obtained by growth transformation of regression functions spanning a Reproducing Kernel Hilbert Space (RKHS). Experiments on TI-DIGIT demonstrate consistent robustness of the new features to noise of varying statistics, yielding significant improvements in digit recognition accuracy over identical models trained using Mel-scale cepstral features and evaluated at noise levels between 0 and 30 dB SNR.

1. INTRODUCTION

While most current speech recognizers give acceptable recognition accuracy for clean speech, their performance degrades when they are subjected to noise present in practical environments [1]. For example it has been observed that additive white noise severely degrades the performance of Mel-cepstra based recognition systems [1, 2]. This performance degradation is attributed primarily to unavoidable mismatch between training and recognition conditions. To reduce the effect of mismatch several techniques have been proposed in the literature, which can be broadly categorized as:

- Noise estimation and filtering that reconditions the speech signal based on noise characteristics [2];
- On-line model adaptation to reduce the effect of mismatch in training and test environments [3];
- Extraction of speech features robust to noise [4], including features based on human auditory modeling [5, 6].

This paper proposes a novel feature extraction mechanism for speech signal representation, *kernel predictive coding cepstra (KPCC)*, by growth transformation on functionals in a reproducing kernel Hilbert space (RKHS). RKHS techniques have been used in signal processing for the purpose of signal estimation and detection in the form of covariance functionals [7]. Our work is based on regression techniques using RKHS which are popular in the machine learning community especially in the field of regularization theory [8] and support vector machines [9]. By imposing smoothness constraints on the functions in RKHS, a nonlinear regression can then be performed to filter out the noise in the signal. It has been shown [8] that for a specific (Toeplitz) form of the kernel the smoothness constraints correspond to low pass spatial filtering.

Growth transformation is an iterative optimization procedure of homogeneous polynomial cost functions constrained over fixed manifolds [10]. This technique is popular for its use in discriminative hidden Markov model (HMM) training using maximum mutual information (MMI) [11], where it is extended to optimizing non-homogeneous rational functions. For this feature extraction task the growth transformation is defined over a parameterized polynomial kernel which over a fixed manifold results in nonlinear features that are very robust to noise and interference.

The paper is organized as follows. Section 2 introduces notions of kernel based predictive coding and growth transformation. Section 3 details the feature extraction algorithm and its parameterization. Section 4 presents results from recognition experiments performed with the resulting features, and Section 5 provides concluding remarks, discussions and future directions.

2. RKHS PREDICTIVE CODING

This section reviews fundamentals and fixes notational convention in light of the KPCC feature extraction algorithm. The key components illustrated in Figure 1 are the kernel regression block and growth transformation block.

This work was supported by a grant from the Catalyst Foundation, New York.

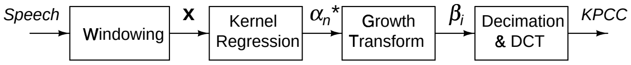

Fig. 1. *Signal flow in KPCC feature extraction*

2.1. Kernel Regression

Given a stationary discrete time signal $x[n], n \in 1, .., N$ and a RKHS H defined over some domain $\Omega \subset \Re^P$, the aim of kernel regression is to estimate a functional $f \in H$ such as to reconstruct (predict) the signal $x[n]$ from previous P samples $\mathbf{x}[n-1] = [x[n-1], .., x[n-P-1]]^T$ as $\hat{x}[n] = \langle f, K(., \mathbf{x}[n-1]) \rangle_H$. Here $\langle ., . \rangle_H$ defines an inner-product between two elements of H and $K : \Re^P \times \Re^P \rightarrow \Re$ is a reproducing positive-definite kernel over H. The optimum function f is obtained by minimizing the cost function

$$\min_{f \in H} C(f) = \frac{\lambda}{2} \|f\|_H^2 + \sum_{n=1}^{N-P} L(x[P+n] - \hat{x}[P+n]) \tag{1}$$

where $L(.) \geq 0$ is a loss function penalizing the reconstruction error, and the smoothness term $\|f\|_H^2$ represents the regularizer penalizing large signal excursions, weighted by regularization parameter λ. The solution of the optimization problem (1) is well known [12] of the general form $f(\mathbf{y}) = \sum_{n=1}^{N-P} \alpha_n K(\mathbf{x}[P+n], \mathbf{y})$. Denoting $K_{nm} = K(\mathbf{x}[P+n], \mathbf{x}[P+m])$ as the kernel matrix and re-substituting in the cost function (1) for square loss leads to ridge regression with dual formulation

$$W(\alpha; \mathbf{K}) = 1/2\lambda \alpha^T (\mathbf{K} + \lambda I)\alpha - \alpha^T \mathbf{x}[N] \tag{2}$$

and with optimum solution

$$\alpha^* = \lambda(\lambda I + \mathbf{K})^{-1} \mathbf{x}[N] \,. \tag{3}$$

.

2.2. Kernel Growth Transformation

The principle of growth transformation can be directly applied to parameterization of an inner-product based kernel. Consider a kernel of the form $K(\mathbf{x}, \mathbf{y}) = g(\langle \mathbf{x}, \mathbf{y} \rangle)$, where $\langle \mathbf{x}, \mathbf{y} \rangle$ is an inner-product defined on two vectors in $\Re^P$, and $g(.)$ is an arbitrary function decomposable as polynomial expansions, like $g(x) = x^2$ or $g(x) = \exp(x)$. The idea behind growth transformation of the kernel $K(., .)$ is to parameterize the inner-product $\langle ., . \rangle$ with predictive coefficients $\beta = [\beta_i]^T, \beta_i \geq 0$ such that

$$\langle \mathbf{x}[n], \mathbf{x}[m] \rangle = \sum_{i=1}^{P} \beta_i x[n-i] x[m-i] + \gamma \tag{4}$$

In addition we enforce that β_i lie on the manifold $M : \sum_i \beta_i = 1$ to ensure proper normalization.

The parameterization in β_i endows the kernel with the following properties:

- The kernel function $K(., .)$ and hence the dual cost $W(\alpha; \mathbf{K})$ is polynomial in β.

- The kernel contains higher-order correlation terms of the discrete signal $x[n]$ and its delayed versions. In this sense the kernel expansion is similar to linear predictive coding (LPC), where the coefficients β_i weigh the correlation across samples at time lags i, although the relationship is nonlinear through the map $g(\cdot)$ in the kernel.

The polynomial nature of dual (2) supports direct application of growth transformations with respect to the parameters β_i to maximize the cost function over the manifold M [10]. The idea behind the growth transformation is to find coefficients β_i to maximally un-learn the regression function $f(\mathbf{x})$ by maximizing the dual (2) obtained by minimization over α. Maximimization over the coefficients β_i yields a transformation emphasizing dimensions in the input vector $\mathbf{x}$ less sensitive to variability, thereby producing a robust kernel that is less prone to noise in the input. Using the growth transformations procedure described in [11] over the manifold $M : \sum_i \beta_i = 0, \beta_i \geq 0$ the coefficients β are re-mapped according to

$$\beta_i \leftarrow \frac{\beta_i \partial W(\alpha^*; \mathbf{K})/\partial \beta_i + D}{\sum_k (\beta_k \partial W(\alpha^*; \mathbf{K})/\partial \beta_k + D)} \,. \tag{5}$$

The parameter D is a smoothing constant that determines the degree of deviation of the new parameters with respect to the old parameters and plays an important role in noise robustness.

Insight into the role of the transformation (5) can be gained by noting that

$$\frac{\partial W(\alpha^*; \mathbf{K})}{\partial \beta_i} = \sum_{n,m} \alpha_n^* \alpha_m^* K'_{ij} x[n-i] x[m-i] \tag{6}$$

$$= \left\| \sum_m \alpha_m^* x[m-i] \Phi'(\mathbf{X}[m]) \right\|_{H'}^2 \tag{7}$$

where $\mathbf{K}'$ is the derivative of kernel $\mathbf{K}$. For the kernels under consideration, $\mathbf{K}'$ also has a reproducing property over a corresponding Hilbert space H'. For the specific choice $K(\mathbf{x}, \mathbf{y}) = \exp(\langle \mathbf{x}, \mathbf{y} \rangle)$, both RKHS representations coincide $H' = H$ and thus the coefficients β_i in (7) reduce to relative norms of the functions in H re-weighted by the training samples at time lags i. This in turn reduces to cross-correlation between the regression function and samples $x[m-i], \forall m$. As an interesting limiting case, for

$\lambda \rightarrow \infty$ and for a linear kernel of type $K(\mathbf{x}[n], \mathbf{x}[m]) = (\sum_{i=1}^{P} \beta_i x[n-i]x[m-i] + \gamma)$, according to (5) the coefficients $\beta_i \propto (\sum_m x[m]x[m-i])^2$ relate directly to the L_2 norm of the auto-correlation function. This affirms that the transformation (5) subsumes linear predictive coding and produces more general, nonlinear features.

3. KPCC FEATURE EXTRACTION ALGORITHM

The KPCC feature extraction procedure illustrated in Figure 1 comprises the following steps:

- Extract speech samples by shifting a rectangular window of size N by W intervals;
- For a given order P choose an initial value of parameters $\beta_i, i = 1, .., P$. In the experiments below the initial values were chosen according to the profile $\beta_i = c + h \sin(i\pi/P)$, akin to the liftering profile in Mel-scale filterbank cepstral coefficient (MFCC) feature extraction;
- Obtain the kernel matrix K by applying the map $g(.)$ to (4) over the data window. Train the dual objective (2) by assigning optimal coefficients α_n^* according to (3);
- Perform growth transformation (5) to obtain new estimates of β_i, at the optimum value of α_n^*;
- Average and decimate the coefficients β_i along $i = 1, \ldots P$ to reduce the number of features; and
- Perform discrete-cosine transformation (DCT) on the reduced coefficients to obtain the final KPCC features. As in MFCC feature extraction, the first DCT coefficient and higher-order coefficients are discarded since they carry little information relevant to speech.

4. EXPERIMENTS AND RESULTS

For all experiments KPCC features were extracted using a 20 ms window shifted by 10 ms, with kernel regression order $P = 60$, and with $c = 0.3, h = 0.5, \lambda = 0.5, D = 1$ and $\gamma = 0.3$. The 60 growth features β_i were averaged and decimated to 30 coefficients. Without loss of generalization it has been assumed that the input signal is rescaled such that $x[n] \leq 1, \forall n$. After DCT, 12 coefficients (indices 2 through 13) were selected as features for the recognition system. Figure 2 shows a sample comparison between KPCC features and corresponding MFCC features for digit *five* obtained before DCT operation for different SNR levels. As standard in MFCC, a window size of 25 ms with an overlap of 10 ms was chosen, and cepstral features were obtained from DCT of log-energy over 24 Mel-scale filter banks. The degradation of spectral features for MFCC in the presence of white noise is evident, whereas KPCC features prevail at elevated noise levels.

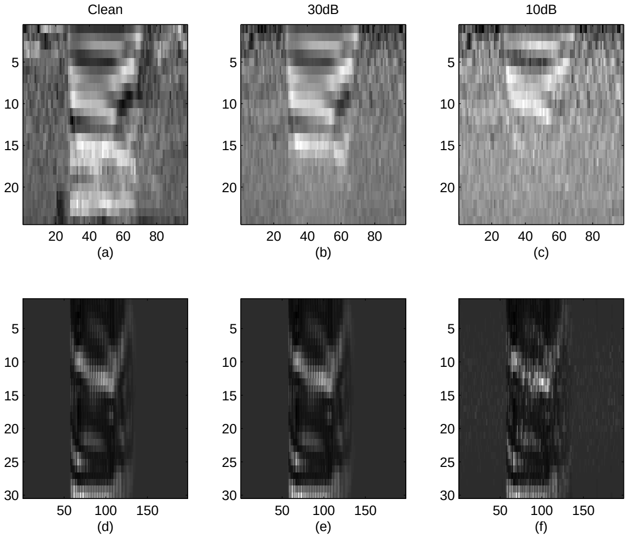

Fig. 2. *MFCC features (a)-(c) and KPCC features (d)-(f) for digit* five *obtained before DCT, under different SNR conditions (clean, 30 dB and 10 dB).*

For recognition experiments, we chose a simple isolated TI-DIGIT digit recognition task with a vocabulary size of 11 (zero to ten plus ‘O’). The training set contained two utterances of isolated digits each from 35 male speakers comprising a total of 770 utterances, and the test set contained isolated digits from 25 other male speakers for a total of 440 utterances. A recognition system was developed using the Hidden Markov Toolkit (HTK), implementing a 14-state left-to-right transition model for each digit where the probability distribution on each state was modeled as a four-mixture Gaussian. As a baseline, the same recognition system was developed using MFCC features comprising of 12 coefficients, without energy and delta features. Noise samples for the experiments were obtained from the $NOISEX$ database and were added to clean speech to obtain test data. We considered four types of noise common in application environments: white noise (W), speech babble noise (B), factory noise (F, plate-cutting and electrical welding equipment) and car interior noise (C, Volvo 340 at 75 mph under rainy conditions). Table 1 summarizes the recognition rates obtained based on the two features under different noise statistics and under different SNR levels.

The following can be inferred from the tabulated results:

1. For clean speech the performance of both systems are comparable, with high recognition rates.
2. For white noise the recognition system with KPCC features demonstrates much better noise robustness

Table 1. Comparison of recognition rates for an HMM system using KPCC features with identical system using MFCC features, for additive white gaussian (W), babble (B), car (C) and factory (F) noise, at various SNR levels.

		Clean	30dB	20dB	10dB	0dB
W	MFCC	98.8%	81.1%	27.5%	12.2%	12.3%
	KPCC	97.8%	97.5%	96.3%	90.6%	48.6%
B	MFCC	98.8%	97.2%	93.8%	60.7%	54.2%
	KPCC	97.8%	96.6%	95.2%	78.4%	54.3%
C	MFCC	98.8%	98.6%	98.1%	96.8%	-
	KPCC	97.8%	97.7%	96.8%	95.2%	-
F	MFCC	98.8%	95.9%	67.7%	28.6%	-
	KPCC	97.8%	96.3%	96.1%	73.86%	-

than corresponding MFCC features. In fact, KPCC maintains acceptable ($> 90\%$) recognition performance for noise reaching signal levels (SNR approaching 0 dB).

3. KPCC features demonstrate significantly better performance in the presence of factory noise and slightly better performance in the presence of babble noise. An interesting observation can be made at this point by noting the trend in recognition rates for babble noise in comparison with other noise types. Babble noise primarily consists of speech signals produced by other humans and hence not only corrupts the entire information bearing frequency bands but also shares statistical properties of the reference signal. This attribute is reflected by reduction in error rates even though KPCC features are more robust to MFCC features. For other sources of noise the statistics are substantially different from reference statistics, which KPCC features utilize to extract noise robust features. This can be observed especially for white noise at very low SNR, for which KPCC features provide reasonable recognition performance.

4. The performance of both MFCC features and KPCC features do not degrade rapidly in the presence of car noise and yield similar relative decrease in recognition rates. This can be attributed to the very low frequency nature of car noise, which keeps the higher frequency features intact for recognition purposes.

5. CONCLUSIONS

In this paper we presented a novel speech feature extraction procedure robust to noise with different statistics, for deployment with recognition systems operating under a wide variety of conditions. The approach is primarily data driven and effectively extracts nonlinear features of speech that are largely invariant to noise and interference with varying statistics.

Acknowledgments: The authors thank Shankar Kumar and Yongang Deng of CLSP at JHU for their help in developing the baseline MFCC system using HTK.

6. REFERENCES

[1] Gong, Y., “Speech recognition in noisy environments: A survey”, *Speech Communication*, 16:261-291, 1995.

[2] Vaseghi, S.V., Milner, B.P, “Noise-adaptive hidden Markov models based on Wiener filters”, *Proc. European Conf. Speech Technology*, Berlin, 1993, Vol. II, pp.1023-1026.

[3] Nadas, A., Nahamoo, D. and Picheny, M.A, “Speech recognition using noise-adaptive prototypes”, *IEEE Trans. Acoust. Speech Signal Process.* Vol.37, No. 10, pp-1495-1502.

[4] Mansour, D. and Juang, B.H, “The short-time modified coherence representation and its application for noisy speech recognition”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, New York, April 1988.

[5] Ghitza, O., “Auditory nerve representation as a basis for speech processing”, in *Advances in Speech Signal Processing*, ed. by S. Furui and M.M.Sondhi (Marcel Dekker, New York), Chapter 15, pp.453-485.

[6] Hermansky, H., “Perceptual Linear Predictive (PLP) Analysis for Speech”, *The Journal of the Acoustical Society of America*,87:1738-1752, April 1990.

[7] Kailath T., and Weinert, H., “An RKHS Approach to Detection and Estimation Problems – Part II: Gaussian Signal Detection”, *IEEE Trans. Information Theory*, vol. **21**, pp 15-23, January 1975.

[8] Girosi, F., Jones, M. and Poggio, T. “Regularization Theory and Neural Networks Architectures,” *Neural Computation*, vol. **7**, pp 219-269, 1995.

[9] Vapnik, V. *The Nature of Statistical Learning Theory,* New York: Springer-Verlag, 1995.

[10] Baum L.E., and Sell, G., “Growth transformations for functions on manifolds”, *Pacific J. Math.*, vol.27, no.2, pp.211-227, 1968.

[11] Gopalakrishnan P.S, et. al, “An inequality for rational functions with applications to some statistical estimation problems”, *IEEE Trans. on Information Theory*, vol. **37**, January 1991.

[12] Wahba, G. *Spline Models for Observational Data*, CBMF-NSF Regional Conference Series in Applied Mathematics, vol. 59, Philadelphia PA: SIAM, 1990.