EMBEDDING TIME WARPING IN EXEMPLAR-BASED SPARSE REPRESENTATIONS OF SPEECH

Emre Yılmaz, Jort F. Gemmeke, and Hugo Van hamme

Dept. ESAT, KU Leuven, Leuven, Belgium

ABSTRACT

This paper describes a new sparse representation model for speech that allows time warping as an extension to a recently proposed sparse representations-based speech recognition system. This recognition system uses exemplars to model the acoustics which are labeled speech occurrences of different length extracted from the training data. Exemplars are organized in multiple dictionaries on the basis of their class and length. Input speech segments are approximated as a sparse linear combination of the exemplars using these dictionaries and a reconstruction error-based decoding is adopted in order to find the best matching class sequence. With the current sparse representation model using a dictionary and a weight vector to approximate an input speech segment, it is not possible to compare input speech segments with exemplars of different lengths. The goal of this work is to introduce a novel sparse representation model which allows time warping using a third matrix which linearly combines consecutive frames in order to shrink or expand the approximation. Preliminary results have shown the feasibility of the proposed sparse representation model.

Index Terms— Exemplar-based speech recognition, sparse representations, time warping

1. INTRODUCTION

Automatic speech recognition has been dominated by statistical acoustic modeling tools, e.g. Hidden Markov models, for several decades. The success of recently proposed speech recognition systems based on exemplar matching attracted considerable interest in exemplar-based acoustic modeling as a viable alternative [1]. These techniques use real speech data, either called exemplars or templates, to recognize unseen speech. Exemplars are labeled speech segments such as phones, syllables or words, possibly of different length, that are extracted from the training data. Each exemplar is tagged with meta-information including speaker, environmental characteristics and prosodic information. Inconsistent exemplar sequences, e.g. mixed gender exemplar sequences, can be penalized based on the tagged meta-information during recognition. An input speech segment can be classified by evaluating the labels of the closest exemplars obtained using a distance metric.

Although exemplars provide better duration and trajectory modeling compared to Hidden Markov Models, they are poorer in terms of generalizability. To cope with this shortcoming, large amounts of data are required to handle the acoustic variation among different utterances [2]. Furthermore, the acoustic distance between the input speech segments and exemplars is found using the dynamic time warping algorithm (DTW). DTW is a well-known algorithm used for matching frame sequences of different lengths in various applications such as speech recognition [3, 4, 5], image recognition [6], audio classification [7] and data mining [8].

An alternative exemplar-based recognition technique is called exemplar-based sparse representations (SR) in which the spectrogram of input speech segments are modeled as a sparse linear combination of exemplars of the same length. SR-based techniques have been successfully used for speech enhancement [9], feature extraction [10] and clean [11] and noisy [12, 13, 14] speech recognition. We have recently proposed an SR-based speech recognition system which uses exemplars of different length organized in separate dictionaries on the basis of their class and length [15]. Compared to a system using fixed-length exemplars stored in a single dictionary, using separate dictionaries for each class provides better classification as input speech segments are approximated as a linear combination of exemplars belonging to the same class only. We have also shown that this system performs reasonably well under noisy conditions in [16].

According to our knowledge, previous SR-based speech recognition systems do not embody a time warping mechanism that allows the comparison of the different-length segments. This paper proposes a novel sparse representation model of speech that embeds time warping in the previous model consisting of a dictionary and a weight vector. Time warping is achieved by means of a sparsely structured warping matrix that learns weights to linearly combine corresponding frequency bands in consecutive frames. The design of the warping matrix has to be handled carefully as too much flexibility in time warping may lead to unrealistic warping. Therefore, only a few successive frames should be combined to approximate an input speech frame. Moreover, sparsity regularization is imposed on the warping matrix to obtain linear combinations often dominated by a single frequency band. This constraint results in approximations that are close to one of the actual frequency bands rather than random linear combinations.

The proposed system differs from classical DTW in several aspects. One main difference is that the proposed model performs a frequency band-level warping by learning distinct weights for each frequency band in a frame, whereas classical DTW provides a frame-level mapping between the time axes. The proposed warping scheme is expected to be more robust against spectral asynchronies, i.e. channel effects in the form of frequency-dependent delays, as it is able to compensate temporal jitters depending on the number of linearly combined successive frames, e.g. when two successive frames are linearly combined, a spectral asynchrony with temporal jitter of a frame shift (typically 10 ms) can be handled. In this sense, the proposed recognizer better models human hearing which is not sensitive to spectral asynchronies up to 40 ms [17].

The rest of the paper is organized as follows. The proposed sparse representation model allowing time warping is given in Section 2. Section 3 explains the experimental setup and implementation details. In Section 4, we present the recognition results and a discussion on the proposed model and its relations with classical DTW is given. The conclusions and thoughts for future work are discussed in Section 5.

2. SPARSE REPRESENTATION MODEL OF SPEECH WITH TIME WARPING

2.1. Previous Model

In the sparse representation model described in [15], the input segments are modeled as a linear combination of the exemplars that are stored in the dictionaries. Each exemplar represents a certain speech unit and the duration of each speech unit in the training data is preserved resulting in exemplars of different lengths. Exemplars spanning l_e frames are reshaped into a single vector with Fl_e time-frequency cells where F is the number of frequency bands in a frame. These reshaped exemplars are stored in the columns of the dictionary $\mathbf{S}_{c,l_e}$: one for each speech unit c and each length l_e. Each dictionary is of dimensionality $Fl_e \times N_{c,l_e}$ where N_{c,l_e} is the number of available exemplars of class c and length l_e.

The baseline model approximates a reshaped input speech vector $\mathbf{y}_{l_i}$ of length Fl_i as a linear combination of the reshaped exemplars of length Fl_e with non-negative weights for each class c:

$$\mathbf{y}_{l_i} \approx \sum_{m=1}^{N_{c,l_e}} \mathbf{s}_{c,l_e}^{m} x_{c,l_e}^{m} = \mathbf{S}_{c,l_e}\mathbf{x}_{c,l_e} \quad \text{s.t.} \quad x_{c,l_e}^{m} \geq 0 \tag{1}$$

where $l_i = l_e$ and $\mathbf{x}_{c,l_e}$ is an N_{c,l_e}-dimensional sparse weight vector. Sparsity of the weight matrix implies that the input speech is approximated by a small number of exemplars. The exemplar weights are obtained by minimizing the cost function,

$$d(\mathbf{y}_{l_i}, \mathbf{S}_{c,l_e}\mathbf{x}_{c,l_e}) + \Lambda \sum_{m=1}^{N_{c,l_e}} x_{c,l_e}^{m} \qquad \text{s.t.} \qquad x_{c,l_e}^{m} \geq 0 \tag{2}$$

where Λ is a scalar which controls how sparse the resulting vector $\mathbf{x}_{c,l_e}$ is. The first term is the divergence between the input speech vector and its approximation. The second term is a regularization term which penalizes the l_1-norm of the weight vector to produce a sparse solution. The generalized Kullback-Leibler divergence (KLD) is used for d:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k \tag{3}$$

The regularized convex optimization problem can be solved using various methods including non-negative sparse coding (NSC). For NSC, the multiplicative update rule to minimize the cost function (2) is derived in [12] and is given by

$$\mathbf{x}_{c,l_e} \leftarrow \mathbf{x}_{c,l_e} \odot (\mathbf{S}_{c,l_e}^{T}(\mathbf{y}_{l_i} \oslash (\mathbf{S}_{c,l_e}\mathbf{x}_{c,l_e}))) \oslash (\mathbf{S}_{c,l_e}^{T}\mathbf{1} + \Lambda) \tag{4}$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}$ is a Fl_e-dimensional vector with all elements equal to unity.

2.2. Proposed Model

To be able to generalize the approximation in Equation (1) to input speech segments of length l_i for $l_i \neq l_e$, we introduce a sparse warping matrix $\mathbf{D}_{c,l_i,l_e}$ of dimensionality $Fl_i \times Fl_e$. For the sake of conciseness, we use $\mathbf{D}$, $\mathbf{S}$, $\mathbf{x}$ and N to represent $\mathbf{D}_{c,l_i,l_e}$, $\mathbf{S}_{c,l_e}$, $\mathbf{x}_{c,l_e}$ and N_{c,l_e} respectively. This warping matrix linearly combines the successive frames to shrink or expand the approximation $\hat{\mathbf{y}}_{l_e} = \mathbf{S}\mathbf{x}$. Thus, a reshaped input speech vector $\mathbf{y}_{l_i}$ can be approximated as a linear combination of the time-frequency cells belonging to successive frames in $\hat{\mathbf{y}}_{l_e}$ for $l_i \neq l_e$,

$$\mathbf{y}_{l_i} \approx \sum_{n=1}^{Fl_e} \mathbf{d}^{n} y_{l_e}^{n} = \mathbf{D}\hat{\mathbf{y}}_{l_e} \tag{5}$$

where $\mathbf{d}^n$ is the n^{th} column of the warping matrix $\mathbf{D}$. Combining Equation (1) and (5), the complete model can be written as

$$\mathbf{y}_{l_i} \approx \sum_{n=1}^{Fl_e} \sum_{m=1}^{N} \mathbf{d}^{n} s^{n,m} x^{m} = \mathbf{DSx} \quad \text{s.t.} \quad x^{m} \geq 0. \tag{6}$$

The new cost function is comprised of three components,

$$d(\mathbf{y}_{l_i}, \mathbf{DSx}) + \Lambda \sum_{m=1}^{N} x^{m} + \beta \sum_{n=1}^{Fl_i} \sum_{m=1}^{Fl_e} d^{n,m} \quad \text{s.t.} \quad x^{m} \geq 0 \tag{7}$$

where β is a scalar which control how sparse the resulting warping matrix is. In this cost function, there is a second regularization term which penalizes the l_1-norm of the rows of the warping matrix to induce sparsity. It should be noted that the structural sparsity of the warping matrix limits the freedom in time warping by allowing only a few consecutive frames with nonzero weights, whereas the regularized sparsity implies that the linear approximation is dominated by a single time-frequency cell obtaining a much larger weight compared to the others. To minimize the cost function in Equation (7), the multiplicative update rules given below are applied iteratively,

$$\mathbf{x} \leftarrow \mathbf{x} \odot ((\mathbf{DS})^{T}(\mathbf{y}_{l_i} \oslash \mathbf{DSx})) \oslash ((\mathbf{DS})^{T}\mathbf{1}_x + \Lambda) \tag{8}$$

$$\mathbf{D} \leftarrow \mathbf{D} \odot ((\mathbf{y}_{l_i} \oslash \mathbf{DSx})(\mathbf{Sx})^{T}) \oslash (\mathbf{1}_D(\mathbf{Sx})^{T} + \beta) \tag{9}$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}_x$ is a Fl_e-dimensional vector and $\mathbf{1}_D$ is a Fl_i-dimensional vector with all elements equal to unity. After each iteration, the rows of the warping matrix $\mathbf{D}$ are normalized to unity in order to avoid extremely small or large values in $\mathbf{D}$ and $\mathbf{x}$. Applying these update rules iteratively, $\mathbf{D}$ and $\mathbf{x}$ become sparser and the reconstruction error between the input speech vector and its approximation decreases monotonically. A reconstruction error-based decoding is applied to find the best matching class sequence using dynamic programming. A known problem of sparse representation approaches working on magnitude spectra is that the silence exemplars are not recognized [12]. This is due to the fact that silence is well-approximated by combining speech exemplars with small weights, so all classes will score equally well. To overcome this problem, reconstruction errors for the class representing silence have to be compensated. The details of the reconstruction error-based decoding and silence dictionary scoring can be found in [15].

2.3. Designing the Warping Matrix

A warping function is defined as a mapping between the time axes of two different patterns (exemplars and input speech segments in this case) [3]. Such a function is expected to capture the spectral similarities between two frame sequences with different durations. To prevent unnatural mappings, some conditions are imposed on the warping function. The warping matrix discussed in Section 2.2 should be properly designed so that it also satisfies these warping function conditions, namely monotonicity, continuity, boundary, adjustment window and slope constraint conditions, which are defined in [3]. Monotonicity and continuity conditions prohibit warping backwards and limit the number of skipped or stalled frames for two consecutive input speech frames. Boundary condition implies matching the first and last frame with the first and last input speech frame respectively. The adjustment window constraint and slope constraint conditions aim to confine the warping path by preventing too many successive skips or stalls.

A warping matrix $\mathbf{D}$ of dimensionality $Fl_i \times Fl_e$ linearly combines the corresponding time-frequency cells belonging to consecutive frames in $\hat{\mathbf{y}}_{l_e}$ to approximate Fl_i input time-frequency cells. Considering the aforementioned conditions, the initial $\mathbf{D}$ matrix is composed of identity submatrices $\mathbf{I}$ of dimensionality $F \times F$ on the diagonal and either sub- or superdiagonal depending on the sign of $l_i - l_e$. For the case of $l_i = l_e + 1$,

$$\mathbf{D} = \begin{bmatrix} \mathbf{I} & 0 & 0 & \cdots & 0 & 0 \\ \mathbf{I} & \mathbf{I} & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{I} & \mathbf{I} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{I} & 0 \\ 0 & 0 & 0 & \cdots & \mathbf{I} & \mathbf{I} \\ 0 & 0 & 0 & \cdots & 0 & \mathbf{I} \end{bmatrix} \tag{10}$$

and $l_i = l_e - 1$,

$$\mathbf{D} = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \mathbf{I} & \mathbf{I} & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \mathbf{I} & \mathbf{I} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \mathbf{I} & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & \mathbf{I} & \mathbf{I} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \mathbf{I} \end{bmatrix} \tag{11}$$

The design can be generalized to any l_i and l_e, once the warping matrix satisfies the warping conditions.

3. EXPERIMENTAL SETUP

3.1. Database

The exemplars used in experiments are speech segments extracted from the clean training set of AURORA-2 database [18] which contains 8440 utterances with one to seven digits in American English. There are 4 clean test sets, each containing 1001 utterances and recognition experiments are performed on these test sets.

3.2. Baseline System

Exemplars and input speech segments are represented in root-compressed (with magnitude power = 0.66) mel-scaled magnitude spectra. A 17 channel mel-scaled filter bank with triangular magnitude response is computed from a spectral analysis with a frame length of 32 ms and a frame shift of 10 ms. The first channel is centered at 200 Hz and the last is at 3030 Hz.

The training data is segmented into the exemplars representing half-digits by a conventional HMM-based recognizer. The system uses 508 dictionaries belonging to 23 different classes. The minimum and maximum exemplar lengths are 5 and 30 frames respectively. Exemplars longer than 30 frames are removed to limit the number of dictionaries. The baseline system uses 10,362 exemplars in total including 260 silence exemplars. Λ is set to 2. The l_2-norm of each dictionary column and reshaped input speech vectors are normalized to unity. Reconstruction error shows enough discrimination among different classes after 50 iterations. Further details about the baseline system can be found in [15].

Table 1. Average word error rates obtained on four clean test sets (SR: Sparse representations, TW: Time warping)

	WER (%)
SR (baseline)	1.91
SR + TW	1.78
SR + TW + Sparsity ($\beta = 10$)	1.66
SR + TW + Sparsity ($\beta = 20$)	1.64
SR + TW + Sparsity ($\beta = 100$)	1.66

3.3. Implementation Details

The proposed system is implemented in MATLAB and GPUs are used to accelerate the evaluation of Equation (8) and (9). We have not made the effort yet to design a dedicated implementation exploiting the sparse structure of the warping matrix $\mathbf{D}$, i.e. in our current implementation the zero entries in $\mathbf{D}$ are reestimated as well. Avoiding this is expected to reduce the simulation times significantly, but requires a significant software engineering effort on a GPU, which has not been performed to date.

4. RESULTS AND DISCUSSION

This section presents the preliminary recognition results obtained using the proposed sparse representation model with time warping. The experiments put more focus on the impact of sparsity regularization imposed on the warping matrix rather than the relative performance of different warping matrix designs. The recognition is performed by approximating input speech segments of length l_i by linearly combining the exemplars of length $l_e = l_i, l_i \pm 1$ using the warping matrices discussed in Section 2.3. These warping matrices linearly combine time-frequency cells belonging to two successive frames to approximate input speech frames except for the first and last input speech frames.

The baseline system uses the sparse representation model described in [15]. The WER obtained with the baseline system is 1.91% which is given in the first row of Table 1. The average simulation time for the baseline system is approximately 3 seconds/utterance. The proposed model with $\beta = 0$ performs better than the baseline with a WER of 1.78% given in the second row. This improvement comes with a great increase in the average simulation time mostly due to the higher number of matrix multiplication in the multiplicative update rules given in Equation (8) and (9). Recognition of each utterance using the proposed model takes 45 seconds on average. After setting β to several nonzero values, $\beta = 10, 20$ and 100 in this case, the WER further reduces to 1.64% for $\beta = 20$. This result shows the positive impact of imposing sparsity regularization on the warping matrix combined with the structural sparsity. This is due to the fact that one of the two time-frequency cells in the consecutive frames gets a much higher weight than the other resulting in a realistic approximation of the input time-frequency cell. Furthermore, it is evident that the recognition accuracy does not vary significantly for different β values. The results discussed above prove the feasibility of the proposed model providing 14%

relative improvement in the WER with time warping limited to a single frame.

The time warping technique we have proposed is different from classical DTW in several aspects. The main difference is that the proposed time warping scheme learns distinct weights for each time-frequency cell whereas classical DTW provides a frame-level mapping between the time axes. One way of adopting a frame-level mapping in the proposed framework is to tie the time-frequency cell weights which belong to the same frame, a constraint for which new multiplicative update formulae have been derived and which will be evaluated in our future work.

Another difference is that classical DTW applies dynamic programming to obtain a warping path through the time axes of the different-length segments. In our case, the complete warping path is learned by fitting a product of matrices to the data. Finally, the conditions on the warping function are imposed more explicitly in classical DTW compared to the proposed approach. The only way to impose these conditions in the proposed scheme is the careful design of the warping matrix. Even with a carefully designed warping matrix, it is not possible to implement some slope constraints such as Itakura constraint [19].

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced a novel sparse representation model for speech signals which allows time warping. This model approximates input speech segments as a product of three matrices, i.e. a sparsely structured warping matrix that linearly combines the time-frequency cells of consecutive frames, a dictionary containing exemplars that are extracted from training data and a weight vector storing the exemplar weights. The design of the warping matrix is of great importance to obtain realistic warping paths. Two warping matrices are introduced for matching two frame sequences with a single frame difference.

Applying this model to recognize digit sequences, we analyze the impact of inducing sparsity in the warping matrix by penalizing the l_1-norm of the rows of the warping matrix. The results have shown that the proposed sparse representation model allowing time warping provides 7% relative improvement in the WER compared to a baseline system which compares input speech segments and exemplars of the same length only. Moreover, the existence of sparsity regularization improves the recognition further yielding a total relative improvement of 14%. This improvements come with a cost of higher computational complexity increasing the average recognition time by a factor of 15, though this number should be interpreted with care given the current sub-optimal implementation.

Even though this preliminary work proved the feasibility of the proposed model, there are still many open questions such as the different warping matrix designs and their effects on the recognition accuracy, a detailed analysis of the effect of different sparsity factors on the recognition accuracy, tying the weights of the time-frequency cells belonging to the same frame to obtain a frame-level time warping and designing a dedicated implementation of the proposed model which is expected to reduce the simulation times.

6. ACKNOWLEDGEMENTS

This work has been supported by the KU Leuven research grant OT/09/028 (VASI) and IWT-SBO Project 100049 (ALADIN).

7. REFERENCES

[1] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98–113, Nov. 2012.

[2] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1377 –1390, May 2007.

[3] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[4] M. De Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Data driven exemplar based continuous speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, 2003, pp. 1133–1136.

[5] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition," in *International Conference on Spoken Language Processing (ICSLP)*, 2006, pp. 2570–2573.

[6] C. A. Glasbey and K. V. Mardia, "A review of image-warping methods," *Journal of Applied Statistics*, vol. 25, no. 2, pp. 155–171, 1998.

[7] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Recognition of isolated musical patterns using context dependent dynamic time warping," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 175–183, May 2003.

[8] D.J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Workshop on Knowledge Discovery in Databases (KDD)*, 1994, pp. 359–370.

[9] J. F. Gemmeke, T. Virtanen, and A. Hurmaleinen, "Exemplar-based speech enhancement and its application to noise-robust automatic speech recognition," in *International Workshop on Machine Listening in Multisource Environments*, Sept. 2011, pp. 53–75.

[10] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representations features for speech recognition," in *Proc. INTERSPEECH*, Sept. 2010, pp. 2254–2257.

[11] J. F. Gemmeke, L. ten Bosch, L. Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *Proc. EUSIPCO*, Glasgow, Scotland, August 24–28 2009, pp. 1755–1759.

[12] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067 –2080, Sept. 2011.

[13] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *Proc. ICASSP*, May 2011, pp. 4588 –4591.

[14] Q. F. Tan and S. S. Narayanan, "Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1337 –1346, May 2012.

[15] E. Yılmaz, D. Van Compernolle, and H. Van hamme, "Combining exemplar-based matching and exemplar-based sparse representations of speech," in *Symposium on Machine Learning in Speech and Language Processing (MLSLP)*, Portland, OR, USA, Sept. 2012.

[16] E. Yılmaz, J. F. Gemmeke, D. Van Compernolle, and H. Van hamme, "Noise-robust digit recognition with exemplar-based sparse representations of variable length," in *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sept. 2012.

[17] T. Arai and S. Greenberg, "Speech intelligibility in the presence of cross-channel spectral asynchrony," in *Proc. ICASSP*, May 1998, pp. 933 –936.

[18] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Tutorial and Research Workshop ASR2000*, Sept. 2000, pp. 181–188.

[19] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 67–72, Feb. 1975.