TOWARDS AUTOMATIC SPEECH RECOGNITION IN ADVERSE ENVIRONMENTS

D. Dimitriadis, N. Katsamanis, P. Maragos, G. Papandreou and V. Pitsikalis

National Technical University of Athens, School of ECE, Zografou, Athens 15773, Greece.
Email:[ddim,nkatsam,maragos,gpapan,vpitsik]@cs.ntua.gr

ABSTRACT

Some of our research efforts towards building Automatic Speech Recognition (ASR) systems designed to work in real-world conditions are presented. The methods we propose exhibit improved performance in noisy environments and offer robustness against speaker variability. Advanced nonlinear signal processing techniques, modulation- and chaotic-based, are utilized for auditory feature extraction. The auditory features are complemented with visual speech cues from the speaker's face, in scenarios where a video stream captured by a simple image-capturing device is available. Speaker adaptation is achieved at the signal level by exploiting certain characteristics of the speech signal which depend on the physical properties of the vocal tract. The proposed methods are overall evaluated on noisy and audiovisual speech databases, i.e. AURORA and CUAVE, and compare favorably to conventional speech recognition systems.

1. INTRODUCTION

Despite intense research, Automatic Speech Recognition (ASR) systems do not yet exhibit acceptable performance in many real life environments. This has seriously undermined the role of ASR as a pervasive *Human-Computer Interaction* (HCI) technology and has limited the applicability of speech recognition systems to well-defined applications like dictation and low-to-medium vocabulary transaction processing systems.

These shortcomings of traditional ASR systems have attracted considerable research in the area. Building robust ASR systems is a very active research field and a variety of methods may be applied to improve speech recognition performance under adverse conditions. These methods encompass improvements at various levels of ASR systems, such as robust feature extraction in the acoustic frontend, adaptation of the ASR system to the environment conditions or to the speaker's characteristics, better statistical modeling in the classifier, enhancement of the acoustic channel with additional cues and better language modeling, to name a few.

Along these lines, this paper highlights the main research directions pursued by our group towards building Automatic Speech Recognition (ASR) systems designed to work robustly in adverse conditions. We first describe acoustic feature extraction schemes which better capture the non-linear dynamics of speech through modulation- and/or chaotic-based modeling of the physics underlying the production of speech. We then describe speaker adaptation methods which better match a multi-speaker ASR system to the distinctive characteristics of a new speaker. Finally, our work in enhancing the speech features with visual cues in an Audio-Visual ASR (AV-ASR) system is presented. Quantitative results on speech recognition in noisy and audiovisual speech databases demonstrate the improved performance of the proposed methods.

2. ROBUST NON-LINEAR ACOUSTIC FRONTEND

2.1. Robust AM-FM Features for Speech Recognition

Robust ASR is an active research field and a variety of algorithms can be used to improve speech recognition performance under adverse conditions including speech enhancement techniques, robust feature extraction and model compensation. In this work, we focus on robust feature extraction schemes.

Motivated by strong evidence for the existence of amplitude and frequency (AM-FM) modulations in speech signals [30], a speech resonance can be modeled by an AM-FM signal,

$$r_i(t) = a_i(t) \cos\left(2\pi \int_0^t f_i(\tau) d\tau\right) \tag{1}$$

and correspondingly the total speech signal as a superposition of a small number of such AM-FM signals (one for each formant). Most often, the number of observable formants does not exheed the 6, and henceforth the speech sounds will be modeled by 6 such AM-FM signals. The estimation of their instantaneous frequencies $f_i(t)$ and amplitude envelopes $|a_i(t)|$ is referred to as the 'Demodulation Problem' and is significant for speech applications.

2.1.1. *Multi-band Demodulation and Feature Extraction*

The AM-FM model suggests the decomposition of speech signals into a series of a few instantaneous frequency and amplitude signals. These signals can be considered as time-frequency distributions containing acoustic information that is not visible in the linear part of the speech spectrum. In [12] preliminary ASR results have indicated that significant part of the acoustic information cannot be modeled by the linear source-filter acoustic model and thus, the need for nonlinear features becomes apparent. These features, which are based on either the FM- or the AM-part, provide additional acoustic information. The modulation features have two major advantages compared to the linear MFCC features. They can model the dynamic nature of speech and capture some of its fine-structure and its rapid fluctuations. Second, they appear to be relatively noise resistant and thus yield improved results, especially for speech recognition in noise, when a mismatch in the training and testing conditions is present.

The AM-FM model suggests that the formant frequencies are not constant during a single pitch period but they can vary around a center frequency. These variations are partly captured by the *Frequency Modulation Percentages* (FMP) features defined as $FMP_i = B_i/F_i$ for each speech resonance i, where B_i is the mean bandwidth (a weighted version of the $f_i(t)$-signal deviation [39]) and F_i is the (amplitude) weighted mean frequency value of resonance i. F_i and B_i are estimated as follows from the information signals $a_i(t)$ and $f_i(t)$ $F_i = \frac{\int_0^T f_i(t) a_i^2(t) dt}{\int_0^T a_i^2(t) dt}$, $B_i = \frac{\int_0^T [\dot{a}_i^2(t) + (f_i(t) - F_i)^2 a_i^2(t)] dt}{\int_0^T a_i^2(t) dt}$ where $i = 1, \ldots, 6$ is the speech resonance index and T the time window length. Another frequency-related feature investigated in this work is the short-time weighted mean of the instantaneous frequency signal $f_i(t)$, i.e. the *Instantaneous Frequency Mean* (IF-Mean). The proposed features provide information about the speech formant fine structure taking advantage of the excellent time-resolution of the ESA. Transitional phenomena and instantaneous formant variations are mapped onto these FM features.

Next, we attempt to model the fine structure of the amplitude envelope signal (AM) with the *Mean Instantaneous Amplitude* (IA-Mean) feature set that is defined as the short-time mean of the instantaneous amplitude signal $|a_i(t)|$ for each speech resonance i. The IA-Mean features parametrize the resonance amplitudes and capture part of the nonlinear behaviour of speech, e.g. the modulation pulses appearing within a single pitch period.

2.1.2. *Feature Extraction Algorithm*

Two significant parts of the feature extraction system are the filterbank and the single-band demodulation algorithm. The AM-FM features are computed from the instantaneous frequency and amplitude signals of each speech resonance. To extract the resonance signals $r_i(t)$ a fixed 6-filter mel-spaced Gabor filterbank is used. The Gabor filters are chosen for several reasons listed in [30], including their optimal time-frequency discriminability. The filter placing and bandwidths are dictated by the mel-scale and the need for constant-Q filterbanks. The bandwidth overlap of adjacent filters is fixed and equal to 50%. Once the resonance signals $r_i(t)$ are extracted, they are demodulated and the $f_i(t)$, $|a_i(t)|$ are obtained.

Among the various demodulation approaches to estimate the model parameters of a single resonance, we use the Energy Separation Algorithm (ESA), due to its excellent time resolution and low complexity [30]. This is based on the continuous-time Teager-Kaiser energy operator (TEO) $\Psi = \dot{x}^2 - x\ddot{x}$. The ESA estimates of the instantaneous frequency and amplitude signals are given by $f(t) \approx (1/2\pi)\sqrt{\Psi[\dot{x}(t)]/\Psi[x(t)]}$ and $|a(t)| \approx \Psi[x(t)]/\sqrt{\Psi[\dot{x}(t)]}$. There is also an ESA for discrete-time AM-FM signals [30]. In this work, we use a more robust ESA where the discrete-time signal is expanded over the continuous-time domain and then, the continuous-time ESA is applied upon.

In order to obtain robust AM-FM features it is crucial that the demodulation algorithm can provide smooth and accurate estimates for $f_i(t)$ and $|a_i(t)|$. There are cases when the demodulation algorithms presented above produce estimates that have singularities and spikes which should be eliminated before the feature measurement process. For this purpose a binomial smoothing of the energy signals is done to smooth out the highpass modeling error of the ESA. Also, a post-processing scheme is applied upon the demodulated instantaneous signals that employs a median filter with a short window.

2.1.3. *Recognition Results*

We have applied the proposed features to the Aurora-3 Speech Database (Spanish task). The 'TIMIT+Noise' databases are created by adding babble, white, pink and car noise to the test set of the TIMIT database which is sampled at 16 kHz; the SNR level is set equal to 10 dB. The ASR experiments have been performed using the HMM-based HTK Tools system [55]. Context-independent, 14-state left-right word HMMs were used; each state contains 16 gaussian mixtures. For the TIMIT recognition tasks, the HMM models are 3-state, left-right HMMs with 16 mixtures. The grammar used for both cases is the all-pair, unweighted grammar. Finally, for the TIMIT+Noise cases, the HMM models are trained in the clean speech training set and tested in the noise-corrupted versions of the testing set.

The input vectors are split into two different data streams, one for the standard features MFCC and the other for the modulation-based features. The data streams are as-

Scenario	WM	MM	HM		Aver. Rel.
Features				Average	Improv.
Aurora Frontend (WI007)	92.94	80.31	51.55	74.93	-
MFCC+CMS (Baseline)	93.68	92.73	65.18	83.86	35.62
MFCC+CMS+IA-Mean	93.22	91.35	71.35	85.31	41.40
MFCC+CMS+IF-Mean	90.71	89.52	72.36	84.20	36.98
MFCC+CMS+FMP	94.38	92.46	72.79	86.54	46.31

Table 1. Correct Word Accuracies (%) for Modulation Features on the Aurora-3 (Spanish Task) Database.

Phoneme Accuracy for the TIMIT Tasks (%) for SNR=10 dB							
	TIMIT	NTIMIT	TIMIT+	TIMIT+	TIMIT+	TIMIT+	Aver. Rel.
			Babble	White	Pink	Car	Improv.
MFCC	58.40	42.42	27.71	17.72	18.60	52.75	-
MFCC+IA-Mean	59.61	43.53	39.25	26.03	31.05	56.50	17.62
MFCC+IF-Mean	59.41	43.70	38.56	26.05	32.81	56.75	19.13
MFCC+FMP	59.92	43.69	38.60	26.15	32.84	55.97	18.17

Table 2. Correct Phoneme Accuracies (%) for Modulation Features on the TIMIT Tasks.

sumed independent. The augmented feature vector consists of 57 coefficients, 39 samples for the ‘standard’ features (normalized energy, MFCCs, 1^{st} and 2^{nd} time-derivatives) and 18 for the modulation features (6 coefficients plus their 1^{st} and 2^{nd} time-derivatives). Cepstral Mean Subtraction (CMS) is applied to the standard feature stream only for the Aurora-3 database to combat convolutional mismatches (for the MM-scenario there is microphone mismatch). The frame length is set equal to 30 msec with frame-period equal to 10 msec. The weights of the two independent data streams are optimized on held-out data. In practice, the stream-weight for the AM-FM features decreases with the SNR level, another indication of the robustness of the proposed features. More specifically, for the clean case i.e. the TIMIT task, the stream weights are set $s_1 = 1.00$ and $s_2 = 0.20$ for the MFCCs and the modulation features, correspondingly. For the low SNR cases the stream weights are $s_1 = 1.00$ and $s_2 = 0.50$ or $s_2 = 1.00$ depending on the noise-level. In Table 2.1.3 and 2.1.3 the recognition results are presented for the Aurora-3 and the TIMIT tasks, respectively. By combining MFCCs with AM-FM features we achieve a performance improvement for the clean and especially for noisy conditions. The improvement is larger for the HM-scenario of the Aurora-3 database and the TIMIT+Noise tasks where additive noise is the main source of degradation. On the other hand, for the NTIMIT and the Aurora-3 WM-, MM-scenarios, where the convolutional noise is dominant, the modulation features yield modest results.

Overall, the AM-FM features provide robustness to additive noise tasks but less so for convolutional noise. Relative error rate reduction up to 46% for mismatched noisy conditions is achieved when these features are combined with MFCCs. We have presented strong indications that modulation features can model and classify different phoneme classes better and more efficiently than the classic MFCC features, especially in the presence of additive noise. In our on-going research we are investigating (i) the usefulness of 2^{nd}-order statistics of the modulation signals, and (ii) ways of optimally combining linear and modulation features for ASR tasks.

2.2. Robust Dynamical Processing & Fractal Features for Speech Recognition

2.2.1. Introduction

There has been strong experimental and theoretical evidence for the existence of important nonlinear aerodynamic phenomena in the vocal tract during speech production. Such phenomena [48, 20, 49] include non-laminar flow, flow separation in various regions, generation and propagation of vortices and formation of jets. The above phenomena can lead to the generation of turbulent flow while the air jet may be modulated either by the vibration of the walls or by the generated vortices. It has been conjectured that methods developed in the framework of chaotic dynamics and fractal theory might be employed for the analysis of turbulent flow, for example modeling of the geometrical structures in turbulence (spatial structure, energy cascade) utilizing fractals and multifractals [28, 5].

Numerous methods have been proposed [26, 31, 34, 4, 30, 12] that attempt to exploit turbulence related phenomena of the speech production system [48, 20], that the linear source-filter model cannot take into consideration. Some of them are based on concepts of fractal theory and dynamical systems. Early work in this area includes the application of fractal measures on the analysis of speech signals [29, 31], application of nonlinear oscillator models, [43, 50, 25] generalized fractal dimensions and multifractal analysis [3, 1]. Additional momentum in the field has been introduced by ideas concerning state-space reconstruction. Methods that follow this approach do not make any assumptions about the underlying model and their mathematical background is based on the embedding theorem [45]. Early works in this field are [43, 50, 34, 26], while recent approaches can be found in [4, 23].

We present an alternative denoising scheme from the dynamical systems' perspective for speech processing. The application of such methods is limited in the current literature [4, 17] especially as far as experiments on extended databases are concerned. In the subsections that follow, we employ methods to filter the signal in the reconstructed space [22, 44, 7, 8] based on the assumption that the unfolded signal is closer to the dynamics of the speech production system when compared to the scalar signal. We further measure invariant quantities of the filtered set (correlation dimension) and use them as acoustic features that quantify the underlying complexity. Next, we evaluate fractal and modulation related features independently on selected parts of the Aurora 2 database. Finally, we merge the twofold nonlinear information in a hybrid feature set and observe that the combination of features relevant to the fractal and the modulation structure of speech can provide additional acoustic information and increased robustness against noise when compared to the typical MFCC features.

2.2.2. Robust Embedding - Fractal Dimensions

We assume that the speech production system may be viewed as a nonlinear dynamical system $X(n) \rightarrow F[X(n)] = X(n+1)$. A speech signal segment $s(n)$, $n = 1, ..., N$, is considered a 1D projection of a vector function applied to the unknown *multidimensional* state variables $X(n)$. A question that naturally arises is whether there exists a reverse procedure by which a phase space $Y = Y(n)$ can be reconstructed - using information provided by the scalar signal - satisfying the requirement to be diffeomorphic to the original phase space, so that determinism and differential information of the dynamical system are preserved.

According to the *embedding* theorem [45] the vector : $Y(n) = [s(n), s(n+T_D), \dots, s(n+(D_E-1)T_D)]$ formed by samples of the original signal delayed by multiples of a constant time delay T_D defines a motion in a reconstructed D_E-dimensional space that has many common invariants with the original phase space of $X(n)$, like fractal dimensions. Thus, by studying the constructible dynamical system $Y(n) \rightarrow Y(n+1)$ we can uncover useful information about the original unknown dynamical system $X(n) \rightarrow X(n+1)$ provided that the unfolding of the dynamics is successful [21], e.g. the embedding dimension D_E is large enough.

The time delay corresponds to the constant time difference between the neighboring elements of each reconstructed vector. The smaller T_D gets, the more the successive elements get correlated. On the contrary, the greater T_D gets, the more random will the successive elements be and any preexisting 'order' will be lost. To compromise between these two conflicting arguments Average Mutual Information I is estimated for the signal $s(n)$ [21]: $I(T) = \sum_{n=1}^{N-T} P(s(n), s(n+T)) \cdot \log_2 \left[\frac{P(s(n), s(n+T))}{P(s(n)) \cdot P(s(n+T))} \right]$ where $P(\cdot)$ is a probability density function estimated from the histogram of $s(n)$. $I(T)$ equals the mutual information for a pair of observed values $s(n), s(n+T)$. Then, the 'optimum' time delay is selected as : $T_D = \min\{\arg\min_{T \geq 0} \ I(T)\}$. An alternative method utilizes in a similar heuristic way the linear autocorrelation function.

The final step in the embedding procedure is to set the dimension D_E of the reconstructed vectors. As a consequence of the projection, manifolds are folded and different distinct orbits of the dynamics are intersecting. A true vs. false neighbor criterion is formed by comparing the distance between two points S_n, S_j embedded in successive increasing dimensions. If their distance $d_D(S_n, S_j)$ in dimension D is significantly different from their distance $d_{D+1}(S_n, S_j)$ in dimension $D + 1$, i.e. $R^D(S_n, S_j) = (d_{D+1}(S_n, S_j) - d_D(S_n, S_j))/(d_D(S_n, S_j))$ exceeds a threshold (in the range $[10, 15]$). The dimension D at which the percentage of false neighbors goes to zero (or minimized in the existence of noise) is chosen as the embedding dimension D_E. A review of methods for the selection of the embedding parameters can be found in [21].

Denoising of Embedded Speech Signals

Increased interest has appeared in the field of robust phase space reconstruction [22, 44, 7, 8, 10, 4]. The methods developed may be grouped into global or local modeling techniques. The former are based on the construction of a global approximation model for the whole set. On the other hand, local models process data in the vicinity of local neighborhoods leading to more detailed and possibly more complex models. Another point of discrimination of the different methods is whether they exploit geometrical or dynamical information or both (for a review see [24]).

We consider the clean scalar speech signal $s(k)$ which is contaminated with *additive* noise $\eta(k)$ giving the observed signal $s_n(k) = s(k) + \eta(k)$. The embedded signal $Y_n = Y_n(k)$ will be corrupted by noise and will be characterized by increased variance. Our objective is the suppression of the affected components of the contaminated signal. Hence, the main concept is to apply some sort of transformation function (e.g. smoothing) to reduce the variance of the affected components. This processing shall be applied in the neighborhood $N(k) = \mathcal{N}^m(Y(k))$ of each reference point $Y(k)$ consisting of m points in the multidimensional phase space. Local neighborhoods correspond to points that are close, in the sense of the system's dynamics.

A simple approach considered in [46] is to replace each reference point by the neighborhood's mean $\overline{N(k)}$ and leads to modest results. A more advanced approach that exploits geometrical information of the local neighborhoods

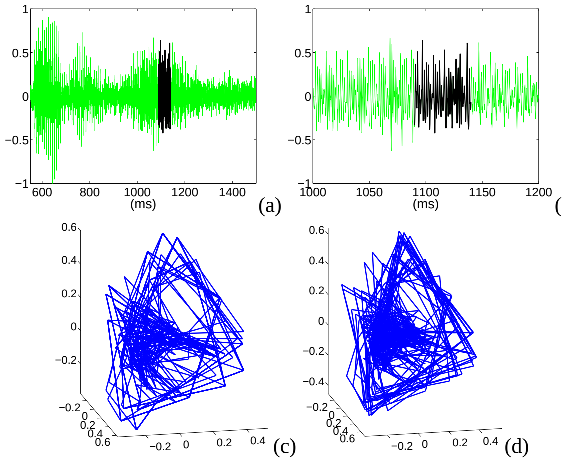


Fig. 1. (a) Noisy speech signal segment of length 950 ms (Aurora 2, SNR: 5 dB); in dark color the 50 ms frame that is processed, magnified in (b). Embedded frame: (c) before and (d) after SVD projective cleaning. Trajectories in (d) are more consistent to the *geometry* of the assumed dynamics, compared to (b).
za

is formed by decomposing the neighborhood data points by Singular Value Decomposition (SVD) [10] onto a set of principal components:

$$N(k) = U \cdot S \cdot V^T, \qquad (2)$$

where U is the matrix formed by eigenvectors of the structure matrix $N(k) \cdot N(k)^T$, S is a diagonal matrix that contains the singular values ordered as $|\sigma_1| \geq \ldots |\sigma_i| \geq \ldots |\sigma_N| \geq 0$ and V is the matrix formed by the eigenvectors of the covariance matrix $N(k)^T \cdot N(k)$.

In the case the noise variance is less than the variance of the signal, the larger eigenvalues will correspond to the system dynamics, whereas the smaller ones to the noise components. To suppress these components we project the data on a subset of principal components $1 \ldots k_p$ that accounts for a percentage λ of the total variance of the local neighborhood set $N(k)$ i.e. $\sum_i^{k_p} \sigma_i^2 = \lambda \sum_i^N \sigma_i^2$. We apply this projection step only on the central reference point and not on the whole neighborhood considered, so as not to distort points that are on the boundaries of the neighborhoods. Additionally, when correcting each reference point we only apply a percentage of the suggested correction and not the whole of it, so as not to risk major corrections in the wrong direction. We adopt this detail [44] in order to move the corrected points incrementally to the positions suggested by the geometry of their neighborhood and not directly at one step, which might introduce instabilities. The projection step described above is applied repetitively i.e. for successive passes over the whole dataset. In case the clean signal is available, a simple ending criterion by computing the SNR gain may be applied. However, no standard approach exists for real data and more advanced techniques should be employed e.g. by utilizing dynamical information [24]. In this preliminary application we adopt a heuristic approach by applying a fixed number of iterations (e.g. from 8 to 12) that we have found to perform satisfactorily in our case. A variation of the procedure of local projections described above on which the results of the next section are based is to apply a lowpass filter [44] on the scalar noisy signal $s_n(k)$ before the embedding procedure. However, the corrections that are computed via the local projections are applied on the original noisy signal. Figure 1 shows a speech segment together with the corresponding embedded frame, before and after applying the projection-based cleaning procedure. One may notice that the trajectories in various regions are more compact, according to the assumed dynamics geometry, compared to the noisy one.

In the unfolded phase space we measure invariant quantities of the dynamically filtered set. *Correlation dimension* can be practically estimated employing a method that belongs to the category of average pointwise mass algorithms for dimension estimation [15]. A quantity that is used for its estimation is the correlation sum C, which is given for each scale r by the number of points with distances less than r normalized by the number of pairs of points: $C(N,r) = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j \neq i} \theta(r - \|X_i - X_j\|)$ where θ is the Heavyside unit-step function. The correlation dimension, which belongs to a superset of generalized dimensions of probabilistic type, is defined as : $D_C = \lim_{r \to 0} \lim_{N \to \infty} \frac{\log C(N,r)}{\log r}$ and for small enough scales and for N large enough $C(r)$ is proportional to r^{D_C}. It corresponds to the number of active degrees of freedom and indicates the underlying system complexity. After estimating C and D_C we create an 8 component feature vector (Filtered embedded Dynamics - Correlation Dimension, FDCD) by including: the mean and the variance of the correlation integral and the mean and the variance of the correlation dimension (D_C) over all scales. We extend the above components by the mean and the variance of D_C over half of the smaller scales $[min(r) \ldots (max(r)+min(r))/2]$ and half of the larger scales $[(max(r)+min(r))/2 \ldots max(r)]$ in order to include an estimate of local (as far as scale is concerned) information on the correlation dimension.

2.2.3. ASR Experiments: Hybrid Fractal-Modulation Features

We extract features based on the nonlinear models presented in the previous section and concatenate them with the typical MFCCs. At first, we present ASR results for the double stream case (i.e. MFCC plus the FDCD) showing relative improvement in the recognition rates and then augment MFCCs with both FDCD and Frequency Modulation Percentages (FMP) presented in the previous section ([12]).

The ASR experiments have been performed on parts of the Aurora 2 [19] database using the HMM-based HTK Tools system. Context-independent, 18-state, left-right word HMMs with 3 gaussian mixtures are used. The grammar used is all-pair and unweighted. Finally, the HMM models are trained in the *clean speech training set* and tested in several noisy test sets.The input vectors are split into different data streams, one for each feature vector and are assumed independent. The augmented features include 13 elements for the 'standard' features (MFCCs plus normalized energy) augmented by 6 feature vector elements for the modulation features and an 8 dimensional feature vector for the fractal features. All feature vectors are extended by their first and second time-derivatives, while Cepstral Mean Subtraction (CMS) is applied to the MFCCs to deal with noise mismatches. The frame length is set equal to 30 ms with frame-period equal to 10 ms; for the fractal features, additional information surrounding each frame is considered by utilizing 50 ms frames. In this way we include both short-time resonance information (FMP) and the complexity information at various scales that may reside in longer frames (FDCD). In Tables 3 and , 4 we present the recognition results for the baseline together with the relative improvements by the augmented feature sets (FDCD). Relative improvement percentage is defined as $(WPA - WPA_{base})/(WPA_{base})$ where WPA is the Word Percent Accuracy using the augmented features. In almost all cases the improvements are increased as the SNR drops (in 5 dB SBR, average improvement of 42.5%). The average improvement over all the presented tests and SNRs is 20.5%. Next, in Table 5 by combining MFCCs with the AM-FM related features (FMP) we achieve a slight performance improvement for higher SNRs and noticeably better improvements for the middle and lower SNRs. Finally, when we combine information from both feature types (FMP + FDCD) (Table 5), we achieve improvements either of the same magnitude as each feature set on its own, or even higher (e.g. SNR of 5 dB). The average improvement for the hybrid augmented features for the presented SNRs is 29.3%.

3. USER ROBUSTNESS

3.1. Introduction

User robustness is a major issue for state of the art automatic speech recognition systems. Their performance may severely degrade due to inter-speaker variations. To tackle this problem, research has moved in two general directions in the last decade: a) Speaker Adaptation, including techniques (Maximum Likelihood Linear Regression, Maximum A Posteriori) that adapt the parameters of the recognition models using speech data by a new speaker and b)Speaker Normalization, that is based on the idea that it is possible to process the new speaker's speech signal in a way that the extracted features better match the Speaker Independent recognition models that have been acquired during the training phase. We have mainly focused on the latter approach.

3.2. Vocal Tract Length Normalization

Probably the most effective method currently used for Speaker Normalization is Vocal Tract Length Normalization. VTLN tries to explicitly compensate for variations in Vocal Tract Length (VTL) among speakers. Variations of the vocal tract shape (especially the VTL) are generally regarded as one of the major sources of inter-speaker variance.

In this context, research on normalization of parametric representations of the speech signal for the purpose of reducing the effects of inter-speaker differences first appeared for vowel identification [14, 51]. Normalization was performed using linear and nonlinear frequency warping functions to compensate for variations in formant positions among speakers. These procedures attempted to solve the difficult problem of estimating the formant positions that correspond to the ”true” vocal tract shape of each speaker, and then compensating for these differences. This problem was avoided by [2] that introduced VTLN in Large Vocabulary Continuous Speech Recognition (LVCSR). A maximum likelihood approach in the ASR framework was suggested to find the best transformation of the signal and a flurry of research activity was sparked. Many variations and improvements of this method have appeared in the literature ever since. Conceptually, three issues are involved in VTLN: 1.Choosing the proper normalization mapping of the speech data, i.e. the proper warping of the spectrum in the frequency axis. 2.Estimation of the speaker-dependent parameters of the mapping, given the speech data, i.e. the warping factor. 3.Application of the normalization mapping.

For the first issue, various warping functions have been proposed and evaluated. Such functions may have the general form: $f' = h(f)$ where f is the original and f' is the new frequency. The warped spectrum is: $Y_h(f) = X(h(f)) = X(f')$. Both linear and nonlinear transformations have been proposed. The method presented in [2] corresponds to a linear frequency warping: $f' = \alpha f$ where α is commonly referred to as the warping factor. Theoretically, this warping can compensate for VTL variations in the cases when the lossless uniform tube model of the vocal tract applies, i.e. for open vowels such as /AA/ [13]. Piecewise linear warping functions that allow for different factors for different parts of the spectrum are commonly used in order to allow the invariance of the bandwidth of the warped spectrum [27].

For the estimation of the warping parameters (i.e. α) given the speech data, various approaches have been used. They may be classified in two broad categories: 1. Fea-

Set A	***Subway***						***Babble***					
SNR (dB)	clean	20	15	10	5	0	clean	20	15	10	5	0
MFCC	98.68	95.64	89.41	71.91	42.27	22.22	98.45	96.01	87.34	64.38	33.61	11.63
+FDCD	-0.13	+0.40	+2.92	+15.38	+43.60	+18.77	+0.06	+1.40	+7.91	+28.91	+69.77	+80.22
Set A	***Car***						***Exhibition***					
SNR (dB)	clean	20	15	10	5	0	clean	20	15	10	5	0
MFCC	98.72	96.31	90.45	72.49	44.48	17.25	98.98	94.72	87.71	66.87	33.40	11.57
+FDCD	-0.12	+0.32	+3.08	+17.29	+34.15	-26.26	-0.24	+1.20	+4.10	+20.76	+71.29	+56.61

Table 3. *Recognition results (Aurora 2/Test A): Baseline (MFCC) and relative improvement percentage*

Set B	***Restaurant***						***Street***					
SNR (dB)	clean	20	15	10	5	0	clean	20	15	10	5	0
MFCC	98.65	95.86	86.72	66.59	38.29	14.13	98.42	96.40	91.07	71.51	44.06	19.79
+FDCD	-0.14	+0.87	+6.82	+21.19	+46.28	+77.21	+0.06	+0.49	+3.45	+17.23	+38.13	+22.54
Set B	***Airport***						***Train-Station***					
SNR (dB)	clean	20	15	10	5	0	clean	20	15	10	5	0
MFCC	98.75	96.10	88.44	69.63	43.30	20.99	99.02	96.59	88.92	71.63	42.47	15.85
+FDCD	-0.07	+0.40	+4.22	+20.42	+40.55	+22.77	-0.23	+0.68	+5.49	+18.18	+41.75	+11.29

Table 4. *Recognition results (Aurora 2/Test B): Baseline (MFCC) and relative improvement percentage .*

Test Set	***Subway***			***Babble***			***Car***			***Exhibition***		
SNR (dB)	20	10	5	20	10	5	20	10	5	20	10	5
MFCC	95.64	71.91	42.27	96.01	64.38	33.61	96.31	72.49	44.48	94.72	66.87	33.40
+FMP	-1.0	+7.4	+30.8	+0.7	+23.8	+61.2	-0.5	+8.4	+23.0	+0.3	+9.6	+43.6
+FMP+FDCD	-0.2	+13.8	+49.6	+1.1	+31.8	+83.9	+0.0	+11.7	+39.8	+1.0	+18.7	+70.7

Table 5. *Recognition results (Aurora 2, selected tests): Baseline (MFCC) and relative improvement percentage .*

ture Based, that exploit properties of the features (e.g. formant locations) to determine the value of the warping factor. 2. Maximum Likelihood Based, that choose the proper factor value from a predetermined set, based on the likelihood of the warped speech data. In [2, 27, 40, 42, 56, 16] the researchers get a Maximum Likelihood (ML) warping factor estimate related to the degree of mismatch between the speaker's warped utterances and the speech recognition Hidden Markov Models. The main concept is that the factor may be given by: $\hat{\alpha} = argmax_{\alpha \in A} P(X_i^{\alpha}|\lambda, W_i)$. X_i^{α} is the speaker's utterance warped in frequency by α , λ is a given HMM set, W_i is a transcription of the utterance either known beforehand (training) or obtained after decoding the unwarped data (testing). In this approach, α is usually chosen from a predetermined set A of values. For the linear warping function for example, α may be chosen from a discrete set A of values between 0.88 and 1.12, which is a range that reflects the 25% range of VTLs of adults [27]. To maintain a stronger link to the VTLN's physiological background and avoid the computational load of possibly multiple decoding passes that the ML approach involves, Eide and Gish [13] estimated the factor based on the position of the 3rd formant in the speaker's utterances.

The way frequency warping is applied to the data may also vary. In [2] the data was properly resampled in the time-domain. Equivalently and more efficiently, Rose et Al [27, 40] warped properly (using the inverse warping factor) the center frequencies of the Mel-Frequency filterbank at the front-end. For example, the speech signal is compressed in the frequency domain if the frequency scale of the filters is stretched. Similarly, if the filterbank frequencies are compressed, the signal frequency scale is stretched. Equivalently, Welling et Al [53], Zhan and Waibel [56], Wegman et Al [52] apply the warping to the speech spectrum just before the filterbank.

Integration of VTLN in an LVCSR system has been reported to work in many different scenarios. In most cases, the speech recognition HMMs are trained using Vocal Tract Length Normalized Utterances. This is straightforward in the approaches for which the estimation of the warping factor is feature-based. Normalization may be independent from training. For the approaches that estimate the warping factor based on ML however, training procedure may be more complex, since the models are involved in the estimation of the factors that are used to normalize the training data.

To cope with this situation, a conceptually simple method is proposed in [53] is . It consists of three steps: 1.An acoustic model consisting of a single Gaussian density per HMM state is trained from the nonnormalized acoustic vectors of all training speakers. 2.For each speaker a warping factor is chosen by ML based on the model trained in the previous step. Forced time alignment is used, since transcriptions of the training utterances are known. 3.The training utterances are normalized and they are used for the training of a normalized model.

For decoding, a warping factor is estimated on a per-utterance basis in most cases. Firstly a hypothesized transcription is obtained using the unwarped data. Then the data is warped by every factor in the predetermined set A and by time alignment, given the hypothesized transcription, the most likely warping factor is chosen. A final decoding pass is performed using the corresponding warped utterance to get the recognition result [27, 57]. An improved two-pass strategy is given by [53] who show that using an unnormalized model to get the preliminary hypothetical transcription may give better results.

3.3. VTLN Implementation

We have implemented a baseline Speaker Normalization setup applying VTLN by utilising the tools provided in the Hidden Markov Model Toolkit (HTK). The setup has been initially implemented for the AURORA 4 Database. In HTK, a piecewise linear warping function is used [16]. Warping is applied in the Mel - filterbank domain as described in [27]. We trained 5-state, 6-mixture triphone-HMMs and used MFCCs along with their derivatives, accelerations and after cepstral mean subtraction. We have experimented with the application of VTLN in the testing phase. We evaluated a supervised scenario for which we estimated per speaker warping factors based on 2 ((adaptation)) utterances and using the ML-criterion as described above. These utterances were chosen as a subset of the one half of the standard set of Aurora 4 testing utterances (164) and we used the other half (166 utterances) for evaluation for the various testing conditions. We have also evaluated VTLN in the testing phase only for Speaker Independent Models trained in multiple (SIM) conditions.

As far as the testing phase is concerned the warping factor for each speaker in the test set, in the supervised scenario is estimated as follows : 1. For $\alpha \in [0.8, 1.2]$sampled every 0.025, testing utterances are warped according to the piecewise linear function of HTK . 2. The warped utterances are aligned by the normalized speech recognition models with the known transcription (supervised scenario). The factor chosen is the one for which the warped data (as a whole) exhibit the maximum likelihood. 3. The speaker's testing utterances are then warped by the per speaker chosen warping factor and decoded. Results are given in Table 6 for the cases of both clean and multicondition training of the Speaker Independent recognition models (CSI, MSI respectively).

Word Error Rate (%)			
Method	***Clean***	***Car***	***Tr. Station***
CSI, No VTLN	13.15	24.75	57.5
CSI + VTLN (2 Utts)	11.60	21.40	53.96
MSI, No VTLN	19.45	16.50	29.65
MSI + VTLN (2 Utts)	17.53	14.66	28.21

Table 6. Results of VTLN (AURORA 4).

4. AUDIO-VISUAL AUTOMATIC SPEECH RECOGNITION

4.1. Introduction

Commercial *Automatic Speech Recognition* (ASR) systems are uni-modal, i.e. they only use features extracted from the audio signal to perform recognition. On the other hand, speech recognition by humans is fundamentally multi-modal. Although audio is the most important source of information for speech recognition, people also use visual cues as a complementary aid in order to successfully perceive speech. The key role of the visual modality is apparent in situations where the audio signal is either unavailable or severely degraded, as is the case with hearing-impaired listeners or very noisy environments, where seeing the speaker's face is indispensable in recognizing what has been spoken. The audiovisual nature of speech recognition is lucidly manifested in well known psychological illusions, such as the *McGurk effect* [33]. These findings provide strong motivation for the Speech Recognition community to do research in exploiting visual information for speech recognition, thus enhancing ASR systems with speechreading capabilities [37, 47].

Research in this relatively new area has shown that multimodal ASR systems can perform better than their audio-only or visual-only counterparts. The first such results where reported back in the early 80's by Petajan [37]. The performance gain becomes more substantial in scenarios where the quality of the audio signal is degraded, as is the case with particularly noisy environments such as a vehicle's cabin [38].

However, the design of robust audio-visual ASR systems, which perform better than their audio-only analogues in all scenarios, poses new research challenges. Two new major issues arise in the design of audio-visual ASR systems [41], namely: (1)*Selection and robust extraction of visual speech features*. From the extremely high data rate of the raw video stream, one has to choose a small number of salient features which have good discriminatory power for speech recognition and can be extracted automatically, robustly and with low computational cost. (2)*Optimal fusion of the audio and visual features*. Inference should be based on the heterogenous pool of audio and visual features in a way that ensures that the combined audiovisual system outperforms its audio-only counterpart in practically all scenarios. This is definitely non-trivial, given that the relative quality of the audio and visual features can vary dramatically during a typical session. In the following, we will briefly describe our ongoing research in the area and how we have tried to address the aforementioned challenges in our audiovisual ASR system.

4.2. Visual Front End Design

The main steps in the processing of the visual modality input are a) the detection and tracking of the speaker's face detection and b) extraction of salient visual features from the *Region Of Interest* (ROI) around the speaker's mouth. These features can be shape-based or appearance-based [18, 41].

In our system, the output of a face detector [54] is used as initial condition for an *Active Appearance Model* (AAM) [9, 32] of faces. AAM have proven particularly effective in modeling human faces for diverse applications, such as face recognition or tracking. In the AAM framework an object's shape is defined by a set of landmark points $\{x_i, i = 1 \ldots N\}$, whose coordinates constitute a shape vector s of length $2N$. We allow for deviations of the shape of objects from a mean face shape s_0 by letting s lie in a low-dimensional manifold. Typically a linear n-dimensional subspace is utilized, yielding: $s = s_0 + \sum_{i=1}^{n} p_i s_i$The deformation of the mean shape s_0 to another shape s defines a mapping of the landmark points. This mapping can be extended to the whole face area by imposing regularity constraints, utilizing e.g. thin plate splines. This procedure yields a wrap $W(x;p)$ mapping each pixel of the face template to a pixel on the exemplar face. The spatial deformation $W(x;p)$ brings the face exemplar I into registration with the face template A.

Analogously, we linearly approximate the face appearance (grayvalued or color) $A(x)$ at point x using a set of "eigenfaces" $\{A_i\}$: $A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x)$, where A_0 is the mean appearance of faces. This allows modeling, among others, lighting variability and person identity.

The shape eigenvectors $\{s_i\}$ in (4.2) and their appearance counterparts $\{A_i\}$ in (4.2) are learned during a training phase, usually using a representative set of hand-labelled face images [9]. The training set shapes are first aligned by means of Procrustes' Analysis and then a PCA of the aligned training set shapes yields the main modes of shape variation $\{s_i\}$. Similarly, the leading principal components of the training set appearance vectors constitute the eigenface set $\{A_i\}$. The first few eigenshapes $\{s_i\}$ and eigenfaces $\{A_i\}$ extracted from the training set we used are depicted in fig. 2.

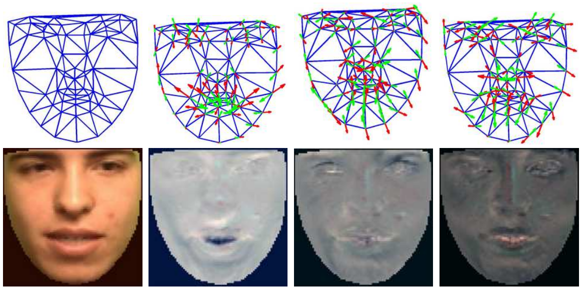
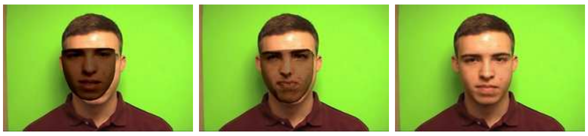


Fig. 2. *Upper row*: Mean shape s_0 and the first eigenshapes s_i. *Lower row*: Mean appearance A_0 and the first eigenfaces A_i.

Model fitting then amounts to finding for each new image the parameters $V \equiv \{p, \lambda\}$ which minimize the appearance reconstruction error in the convex hull of the mean shape (denoted by $x \in s_0$): $V =$

$$\operatorname{argmin}_{p,\lambda} \sum_{x \in s_0} \left[A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) - I(W(x;p)) \right]^2$$

A global similarity transform on the shape and a linear brightness correction on the appearance (not included in eq. (4.2)) are also allowed. Although this is a difficult nonlinear optimization problem if attacked straightforwardly, there are efficient real-time approximate algorithms for iteratively solving it. An example of fitting a face by an AAM with these algorithms is shown in fig. 3. Details on these algorithms are beyond the scope of this paper. One can consult [9, 32]. The fitting procedure uses the output of a face detector as initial shape estimate for the first video frame and it is repeated for each new video frame using the converged solution at the previous frame as starting point. Ultimately a sequence of visual speech features $V_t \equiv \{p_t, \lambda_t\}$ is extracted for each frame t, with $t = 1 \ldots T$.



step 0 step 5 step 15

Fig. 3. Example of the AAM fitting algorithm. As the face mask better localizes the speaker's face, the reconstruction error decreases.

The parameters V of the fitted AAM capture the main modes of variation of both the face's shape and appearance. These parameters, after an optional step of further dimensionality reduction by unsupervised techniques and/or discriminatory training are used for statistical modeling of the visual speech. Inclusion of their time derivatives ΔV and $\Delta\Delta V$, as is usual in the speech recognition community, can be used as an heuristic to capture some of the dynamics of speech and empirically leads to increased recognition performance [41].

4.3. Audio-Visual Feature Fusion

Classical *Hidden Markov Models* (HMMs) do not suffice in modelling the statistics of audiovisual speech. Statistical modelling of multimodal speech is complicated due to issues, such as audio and visual speech asynchrony [6] and varying relative speech discriminative power of the audio and visual streams. Therefore successful audio and visual feature integration requires utilization of advanced techniques and models for cross-modal information fusion.

One can generally classify the various approaches to multimodal feature integration into three main categories [18], depending on the stage that the audio and visual streams are fused, namely early, intermediate and late integration techniques. The early integration paradigm tries to deal with the speech recognition problem utilizing a single classifier (usually a conventional HMM), which acts on the concatenation of the audiovisual features, often after they have undergone an appropriate transformation. The intermediate integration class comprises classification methods that explicitly model the two different modalities and their interaction. The overall class conditional likelihood used in recognition with these models can then be computed by combining the class conditional likelihood of each modality. The inference engines used for these models are usually various HMM extensions, belonging to the general class of *Dynamic Bayesian Networks* (DBNs) [11], which are reviewed in the context of audiovisual speech recognition in [35]. Finally, late integration models utilize different, independent classifiers for the audio and visual features and the final classification decision is reached by combining the partial outputs of the uni-modal classifiers.

4.4. Preliminary Experiments on Audio-Visual Datasets

We have made some preliminary experiments in video-only and audiovisual speech recognition, using the CUAVE audiovisual speech database [36]. The experiments reported here have been conducted in collaboration with A. Potamianos of the Technical University of Crete, Greece.

The CUAVE audiovisual speech database consists of videos of 36 persons, each uttering 50 connected digits. From these sequences, 30 are used for training and the rest 6 for testing. The videos are NTSC-encoded at 29.97 fps. In order to build the visual model, we have hand-labelled the first frame of each video and used it to train the AAM. We have retained 5 shape parameters and 10 appearance parameters ($n = 5$ and $m = 10$ in the notation of 4.2), resulting in a visual feature vector V of length 15. The sequence of visual vectors V_t was upsampled to 100 fps to match the acoustic feature rate and was complemented with the time derivatives ΔV and $\Delta\Delta V$. In our preliminary experiments we have used the simplest form of feature fusion, namely the multi-stream HMM [35] with fixed stream weights, deferring more complete treatment for future work. We experimented with both visual-only and audiovisual scenarios. In the HMM classifier 8-state left-to-right whole digit models were utilized and the observation probabilities were modeled by a single Gaussian pdf. In Table 7 we present word accuracy results on two tasks: digit recognition without endpointing and digit classification, given the ground truth segmentation.

Scenario Features	Audio	Visual	Audio-Visual
Recognition	98	26	78
Classification	99	46	85

Table 7. Correct Word Accuracies (%) for visual/audiovisual experiments on the CUAVE database.

These results highlight the difficulty in fusing the two heterogenous streams, particularly in noiseless scenarios, as in the reported experiment, where the audio stream is far more reliable than the visual stream for speech recognition. Research in the near future will concentrate on the exploration of alternative DBN architectures for feature fusion and on the principled selection of stream weights, depending on the relative reliability of the two streams.

5. ACKNOWLEDGMENTS

Our work has been supported by the European research program HIWIRE, the European Network of Excellence MUSCLE and the NTUA research program ‘Protagoras’.

6. REFERENCES

[1] O. Adeyemi and F. G. Boudreaux-Bartels. Improved accuracy in the singularity spectrum of multifractal chaotic time series. In *Proc. IEEE , ICASSP-97*, Munich, Germany, 1997.

[2] A. Andreou, T. Kamm, and J. Cohen. Experiments in vocal tract normalization. In *CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[3] Y. Ashkenazy. The use of generalized information dimension in measuring fractal dimension of time series. *Physica A*, 271(3-4):427–447, 1999.

[4] M. Banbrook, S. McLaughlin, and I. Mann. Speech characterization and synthesis by nonlinear methods. *IEEE Trans. Speech Audio Processing*, 7:1–17, 1999.

[5] R. Benzi, G. Paladin, G. Parisi, and A. Vulpiani. On the multifractal nature of fully developed turbulence and chaotic system. *J. Phys.*, 18:3521, 1984.

[6] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proc. Int’l Conf. Acoustics, Speech, and Signal Processing*, pages 669–672, 1994.

[7] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson. State space reconstruction in the presence of noise. *Physica D*, 51:52–98, 1991.

[8] R. Cawley and G. H. Hsu. Local-geometric projection method for noise reduction in chaotic maps and flows. *Phys. Rev. A*, 46:3057–3082, 1992.

[9] T.F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proc. Europ. Conf. on Comp. Vision*, volume II, pages 484–498. Springer-Verlag, 1998.

[10] A. G. Darbyshire and D. S. Broomhead. Robust estimation of tangent maps and Lyapunov spectra. *Physica D*, 89:287–305, 1996.

[11] T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Artificial Intelligence*, 93(1-2):1–27, 1989.

[12] D. Dimitriadis and P. Maragos. Robust energy demodulation based on continuous models with application to speech recognition. In *Proc. Eurospeech-03*, Geneva, Switzerland, September 2003.

[13] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 346–348.

[14] G. Fant. Non-uniform vowel normalization. Technical report, Speech Transmiss. Lab., Royal Inst. Technol., Stockholm, Sweden, 1975.

[15] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9:189–208, 1983.

[16] T. Hain, P.C. Woodland, T.R. Niesler, and E.W.D. Whittaker. The 1998 htk system for transcription of conversational telephone speech. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 57–60, 1999.

[17] R. Hegger, H. Kantz, and L. Matassini. Denoising human speech signals using chaoslike features. *Phys. Rev. Lett.*, 84(14):3197–3200, 2000.

[18] M.E. Hennecke, D.G. Stork, and K.V. Prasad. Visionary speech: Looking ahead to practical speechreading systems. In D.G. Stork and M.E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 331–349. Springer, Berlin, Germany, 1996.

[19] H. G. Hirsch and D. Pearce. The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, 2000.

[20] J. F. Kaiser. Some observations on vocal tract operation from a fluid flow point of view. In I. R. Titze and R. C. Scherer, editors, *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, pages 358–386. Denver Center for Performing Arts, Denver, CO, 1983.

[21] H. Kantz and Th. Schreiber. *Nonlinear Time Series Analysis*. Cambridge Univ. Press, Cambridge, UK, 1997.

[22] H. Kantz, Th. Schreiber, I. Hoffmann, Th. Buzug, G. Pfister, L. G. Flepp, J. Simonet, R. Badii, and E. Brun. Nonlinear noise reduction: A case study on experimental data. *Phys. Rev. E*, 48:1529–1538, 1993.

[23] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Trans. Acoust., Speech, Signal Processing*, to apear.

[24] E. J. Kostelich and Th. Schreiber. Noise reduction in chaotic time-series data: A survey of common methods. *Phys. Rev. E*, 48:1752–1762, 1993.

[25] Gernot Kubin. Synthesis and Coding of Continuous Speech with the Nonlinear Oscillator Model. In *Proc. ICASSP'96*, volume 1, page 267, Atlanta, USA, 1996.

[26] A. Kumar and S. K. Mullick. Nonlinear dynamical analysis of speech. *J. Acoust. Soc. Am.*, 100(1):615–629, 1996.

[27] L. Lee and R. Rose. A frequency warping approach to speaker normalization. *IEEE Trans. on Speech and Audio Processing*, 6:49–60, January 1998.

[28] B. Mandelbrot. *The Fractal Geometry of Nature*. Freeman, NY, 1982.

[29] P. Maragos. Fractal aspects of speech signals: Dimension and interpolation. In *Proc. IEEE , ICASSP-91*, 1991.

[30] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. on Signal Processing*, 41(10):3024–3051, October 1993.

[31] P. Maragos and A. Potamianos. Fractal dimensios of speech sounds: Computation and application to automatic speech recognition. *J. Acoust. Soc. Am.*, 105(3):1925–1932, 1999.

[32] I. Matthews and S. Baker. Active appearance models revisited. *Int'l Journal of Comp. Vision*, 60(2):135–164, 2004.

[33] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[34] S. Narayanan and A. Alwan. A nonlinear dynamical systems analysis of fricative consonants. *J. Acoust. Soc. Am.*, 97(4):2511–2524, 1995.

[35] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, (11):1–15, 2002.

[36] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, 2002.

[37] E.D. Petajan. *Automatic Lipreading to Enhance Speech Recognition*. PhD thesis, Univ. of Illinois, Urbana-Campaign, 1984.

[38] R. Pieraccini, K. Dayanidhi, J. Bloom, J.-G. Dahan, M. Phillips, B. Goodman, and K.V. Prasad. Multimodal conversational systems for automobiles. *Communications of the ACM*, 47(1):47–49, January 2004.

[39] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *J. Acoust. Soc. Am.*, 99(6):3795–3806, June 1996.

[40] A. Potamianos and R.C. Rose. On combining frequency warping and spectral shaping in hmm based speech recognition. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, 1997.

[41] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. In G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, editors, *Issues in Visual and Audio-Visual Speech Processing*, chapter 10. MIT Press, 2004.

[42] D. Pye and P.C. Woodland. Experiments in speaker normalisation and adaptation for large vocabulary speech recognition. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 1047–1050.

[43] T. F. Quatieri and E. M. Hofstetter. Short-time signal representation by nonlinear difference equations. In *Proc. IEEE , ICASSP'90, Albuquerque*, April 1990.

[44] T. Sauer. A noise reduction method for signals from nonlinear systems. *Physica D*, 58:193–201, 1992.

[45] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *J. Stat. Phys.*, 65:579–616, 1991.

[46] Th. Schreiber. Extremely simple nonlinear noise-reduction method. *Phys. Rev. E*, 47:2401–2404, 1993.

[47] D.G. Stork and M.E. Hennecke, editors. *Speechreading by Humans and Machines*. Springer, Berlin, Germany, 1996.

[48] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech Production and Speech Modelling W.J. Hardcastle & Marchal, Eds., NATO ASI Series D*, volume 55, 1989.

[49] T. J. Thomas. A finite element model of fluid flow in the vocal tract. *Comput. Speech & Language*, 1:131–151, 1986.

[50] B. Townshend. Nonlinear prediction of speech signals. *IEEE Trans. Acoust., Speech, Signal Processing*, 1990.

[51] H. Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Trans. on Speech and Audio Processing*, 25:183–192, April 1977.

[52] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalization and speaker adaptation, a combination for conversational speech recognition. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, May 1996.

[53] L. Welling, H. Ney, and S. Kanthak. Speaker adaptive modeling by vocal tract normalization. *IEEE Trans. on Speech and Audio Processing*, 10(6):415–426, September 2002.

[54] M. H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58, January 2002.

[55] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Entropic Ltd., 2002.

[56] P. Zhan and A. Waibel. Vocal tract length normalization for large vocabulary continuous speech recognition. Technical report.

[57] P. Zhan and M. Westphal. Speaker normalization based on frequency warping. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 1039–1042.