

Towards Automatic Speech Recognition In Adverse Environments

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

Email: [ddim,nkatsam,maragos,gpapan,vpitsik]@cs.ntua.gr

1. Noise Robustness

1.1. Modulation Analysis

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

$$r_i(t) = A_i \exp\left(\int_0^t b_i(\tau)d\tau\right) \cos\left(2\pi \int_0^t f_i(\tau)d\tau\right) \quad (1)$$

where $b_i(t)$ and $f_i(t)$ are the instantaneous modulating signals. Correspondingly the total speech signal can be considered as a superposition of a small number of such AM-FM signals. The instantaneous Teager energy Ψ of such resonance signal is given by

$$\Psi[r_i(t)] \approx A_i^2 \exp\left(2 \int_0^t b_i(\tau)d\tau\right) 4\pi^2 f_i^2(t) \quad (2)$$

The modulating features based on such signals are proved to be both robust to noise and rich in acoustic information [2]. Herein, we are presenting improved recognition results on noisy tasks when compared to the typical linear MFCC features.

1.2. Dynamical Systems, Fractal Analysis

1.2.1. Phase Space Reconstruction

A speech signal segment $s(n)$ can be thought of as a 1D projection of the unknown *multidimensional* phase space of the speech production system. Possibly, this projection is responsible for a loss of information. By a reverse procedure a multidimensional phase space is reconstructed according to the *embedding* theorem [1]. The reconstructed space $Y(n)$ is formed by samples of the original signal delayed by multiples of a constant time delay T_D and constructs a trajectory in the multidimensional (D_E) space $Y(n) = [s(n), s(n + T_D), \dots, s(n + (D_E - 1)T_D)]$, that shares invariant characteristics with the original phase space. Thus, by studying the reconstructed space we can possibly uncover useful information about the original unknown dynamical system. In the unfolded phase space we measure invariant quantities of the set (attractor) that are conserved from the original phase space. Such measures are the fractal dimensions which correspond to the number of active degrees of freedom and the underlying system complexity.

1.2.2. Robust Processing

Robust processing of speech signals is indispensable for any application that is meant to work in real environments. Denoising in the reconstructed phase space is done in ‘agreement’ to the assumed system dynamics. Common methods for robust processing of speech signals especially in the framework of speech recognition at the feature level, are filterbank analysis, modulation based analysis [2], auditory inspired processing [5],

methods influenced by perceptual ideas [3]. In this work we present an alternative denoising scheme from the dynamical systems’ point of view for speech processing. Based on ideas mentioned above we employ methods to filter the signal on the reconstructed space assuming that the multidimensional signal is closer to the dynamics of the speech production system, compared to the scalar signal.

2. User Robustness

In an effort to achieve user robustness we applied and evaluated state of the art speaker adaptation/normalization methods, like Maximum Likelihood Linear Regression, Maximum A Posteriori Adaptation and Vocal Tract Length Normalization in various scenarios (online, offline, fast adaptation, non-native speech). The results presented exhibit important performance improvement. Our research mainly focused on speaker normalization techniques at the signal level and their proper combination with transformation-based adaptation (MLLR, MAP). Such techniques achieve the proper modification of the incoming utterance based on speaker-dependent properties so that the extracted features may be even more speaker-independent, better suiting to speech recognition models. We tried to apply concepts and algorithms from nonlinear and time-varying speech processing in this direction.

3. Audio-Visual Automatic Speech Recognition

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 [4], namely: 1) *Selection and robust extraction of visual speech features*. 2) *Optimal fusion of the audio and visual features*. In this paper, we will briefly describe our ongoing research in the area and how we have tried to address the aforementioned challenges in our audio-visual ASR system.

4. References

- [1] H. D.I. Abarbanel, *Analysis of Observed Chaotic Data*, Springer-Verlag, New York, 1996.
- [2] D. Dimitriadis, P. Maragos and A. Potamianos, “Robust AM-FM Features for Speech Recognition”, to appear, *IEEE SPL*, 2005.
- [3] H. Hermansky and N. Morgan and A. Bayya and P. Kohn “Compensation for the Effect of the Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP)” *Eurospeech-91*, pp. 578–589, 1991.
- [4] G. Potamianos, C. Neti, J. Luetin, 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.
- [5] S. Seneff, “Pitch and spectral estimation of speech based on an auditory synchrony model”, *ICASSP-84*, pp.3621-3624, 1984.
- [6] D.G. Stork and M.E. Hennecke, editors. *Speechreading by Humans and Machines*. Springer, Berlin, Germany, 1996.