Proceeding of the First International Conference on Modeling, Simulation and Applied Optimization, Sharjah, U.A.E. February 1-3, 2005

A NEW FEATURE EXTRCTION MOTIVATED BY HUMAN EAR

Amin FazelDehkordi
Sharif University of Technology
Computer Engineering
Department, Tehran, IRAN
afazel@mehr.sharif.edu

Hossein Sameti
Sharif University of Technology
Computer Engineering
Department, Tehran, IRAN
sameti@sharif.edu

Sayed Kamal-Aldin Ghiathi
Sharif University of Technology
Computer Engineering
Department, Tehran, IRAN
ghiathi@mehr.sharif.edu

ABSTRACT

Feature extraction method is a very important factor that has great effect on the performance of speech recognition systems. Since living creatures have usually been the best models for human technology, we hope these models can overcome the problem of mismatch between training and testing conditions. In this paper we present a novel feature extraction based on human ear physiology. In this new method, with inspiration of human ear physiology, we model the inner ear for feature extraction. The proposed method is evaluated in a connected digit recognition task. Experimental results show that the proposed feature extraction, called Human Ear based Feature Extraction (HEFE), decreases the relative word error rate by more than 32% with respect to MFCC features.

1. INTRODUCTION

Automatic speech recognition (ASR) is one of the emerging technologies for increasing interaction between human and machine and extracting robust features for automatic speech recognition has been a challenging research domain. Since human speech recognition greatly outperforms current ASR systems in noisy environments, ASR systems seek to improve noise robustness by drawing on physiological inspiration. To build an effective speech recognition system in mimicking human performance, it should guarantee that the processed speech information has the representation close to that of human system. Although most of current methods of feature extraction for speech recognition are based on production model, it seems that methods based on auditory model must be used.

Every recognition task is preceded by an acoustic analysis front-end, aiming to extract significant parameters from the time signal. Normally, this analysis is based on a model of the signal or of the production mechanism. Short-Time Fourier Transform (STFT), Cepstrum, and other related schemes [1] were all developed strictly considering physical phenomena that characterize the speech waveform and are based on the quasi-periodic model of the signal. On the other hand LPC technique and all of its variants were developed directly by modeling the human speech production mechanism. In speech recognition the focus is on perceived sound rather than on physical properties of the signal or of the production mechanism. To this purpose, lately, almost all these analysis schemes have been modified by incorporating, at least at a very general stage, various perceptual related phenomena. Linear prediction on a warped frequency scale STFT-derived auditory models, perceptually based linear predictive analysis, are a few simple examples of how human auditory perceptual behavior is now taken into account while designing new signal representation algorithms. Furthermore, the most significant example of attempting to improve acoustic front-end with perceptual related knowledge, is given by the Mel-frequency cepstrum analysis of speech [2], which transforms the linear frequency domain into a logarithmic one resembling that of human auditory sensation of tone height. In fact, Mel Frequency Cepstrum Coefficients (MFCC) are almost universally used in the speech community to build acoustic front-end for ASR systems. All these sound processing schemes make use of the "short-time" analysis framework [1]. Short segments of sounds are isolated and processed as if they were short segments from a sustained sound with fixed properties. In order to better track dynamical changes of sound properties, these short segments which are called analysis frames, overlap one another. This framework is based on the underlying assumption that, due to the mechanical characteristics of the generator, the properties of the signal change relatively slowly with time. Even if overlapped analysis windows are used, important fine dynamic characteristics of the signal are discarded. Just for that reason, but without solving completely the problem of correctly taking into account the dynamic properties of speech, "velocity"-type parameters and "acceleration"-type parameters [3] have been included in acoustic front end of almost all commercialized ASR systems. The use of these temporal changes in speech spectral representation has given rise to one of the greatest improvements in ASR systems. Moreover, in order to overcome the resolution limitation of the STFT (due to the fact that once the analysis window has been chosen , the time frequency resolution is fixed over the entire time-frequency plane, since the same window is used at all frequencies), a new technique called Wavelet Transform (WT), characterized by the capability of implementing multiresolution analysis, has been introduced[4]. With this new processing scheme, if the analysis is viewed as a filter bank, the time resolution increases with the central frequency of the analysis filters. In other words, different analysis windows are simultaneously considered in order to more closely simulate the frequency response of the human cochlea. As with the preceding processing schemes, this new auditory-based technique, even if it is surely more adequate than STFT analysis to represent a model of human auditory processing, it is still based on a mathematical framework built around a transformation of the signal, from which it tries directly to extrapolate a more realistic perceptual behavior .

Cochlear transformations of acoustic signals result in an auditory neural firing pattern different from the spectral pattern obtained from the waveform by using one of the above mentioned techniques. In other words, spectral representations such as the spectrogram, a popular time-frequency-energy representation of speech, or either the wavelet spectrogram, or scalogram, obtained using the above described multiresolution analysis technique are quite different from the true neurogram. In recent years, basilar membrane, inner cell and nerve fiber behavior have been extensively studied by auditory physiologists and neurophysiologists and knowledge about the human auditory pathway has become more accurate. In this paper we introduce a novel model for basilar membrane and hair cells that make a new transformation on signal and use it for feature extraction.

The remaining paper is arranged as follows: In the next section a brief description of the physiological basis in human auditory system is given. The novel model for the basilar membrane and hair cells is described in the section 3. Section 4 contains the experiment and results followed by conclusions in section 5.

2. PYSIOLOGICAL BASIS IN THE HUMAN AUDITORY SYSTEM

A structure of the human auditory system is shown in Figure 1 [5]. The ear consists of three basic parts: the outer ear, the middle ear, and the inner ear. Each part of the ear serves a specific purpose in the task of detecting and interpreting sound. The outer ear serves to collect and channel sound to the middle ear. The middle ear, which consists of the tympanic membrane and ossicles, serves to transform the energy of a sound wave into the internal vibrations of the bone structure of the middle ear and ultimately transform these vibrations into a compressional wave in the inner ear.

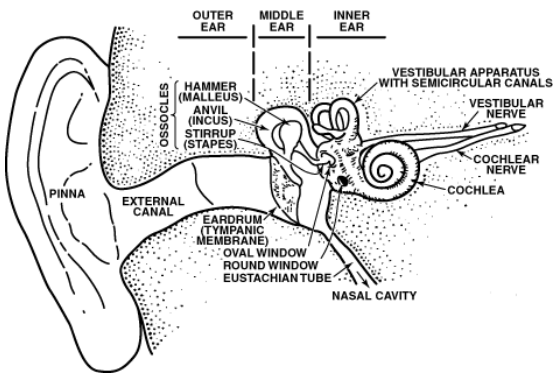

Figure 1: *Structure of human ear.*

The inner ear consists of a cochlea, the semicircular canals, and the auditory nerve. The cochlea and the semicircular canals are filled with a water-like fluid. The fluid and nerve cells of the semicircular canals provide no roll in the task of hearing; they merely serve as accelerometers for detecting accelerated movements and assisting in the task of maintaining balance. The cochlea is a spiral-shaped structure that contains the organ of "Corti", the most important component of hearing. The basilar membrane supports the organ which contains a mass of cells almost touching the branch endings of the auditory nerve. From these cells sprout fine hairs, (23,500 of them) rising in orderly rows like the bristles of a very soft brush. The hairs stick through the dome of the organ, their ends embedded in a thick overhanging sheet, the tectorial membrane. These hairs are transducers. Figure 2 [5] shows basilar membrane and hair cells in the inner ear.

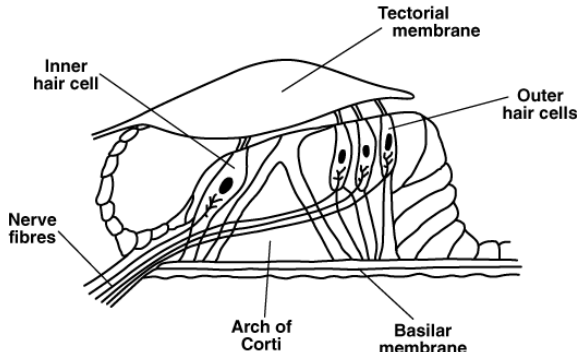

Figure 2: *Basilar membrane and hair cells.*

As the basilar membrane bellies in and out, it pushes and pulls the complex of tissues above it. The hairs' cells of the organ of Corti ride with the basilar membrane. The hairs have their tops embedded in the tectorial membrane and their roots fixed in the hair cells, so the motion of the basilar membrane bends and twists and pulls and pushes the hairs. Each hair cell has a natural sensitivity to a particular frequency of vibration. When the frequency of the compressional wave matches the natural frequency of the nerve cell, that nerve cell will resonate with a larger amplitude of vibration. This increased vibrational amplitude induces the cell to release an electrical impulse which passes along the auditory nerve towards the brain. In a process which is not clearly understood, the brain is capable of interpreting the qualities of the sound upon reception of these electric nerve impulses.

3. MODELLING OF THE BASILAR MEMBRANE AND HAIR CELLS

In this section we explain our simulation of the ear. As shown in figure 3, different parts of basilar membrane and hair cells are sensitive to different frequencies of input signal. As a sinusoid sound with frequency f passes through the basilar membrane, it vibrates those parts of basilar membrane with natural frequency f.

Since corporation of basilar membrane and hair cells changes all frequencies of speech into mechanical energy, with good approximation, we can discretely represent basilar membrane and hair cells as forced damped oscillators with different natural frequencies. We stimulate these oscillators with input sound. At each time we use energy of each oscillator, with natural frequency f, as the energy of input sound at that frequency. Note that we got rid of framing and windowing which are necessary when using FFT.

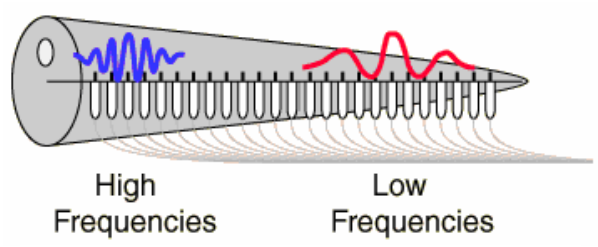

Figure 3: *Our discrete model for basilar membrane and hair cells.*

In this simulation we have an oscillating particle which is always pulled by a force towards the center of oscillation. Displacement of the article from the center of oscillation is shown by x and the inward force is equal to –kx. Where, k is the constant of oscillator. Since we have a foreign force (posed by sound), we can no further use those standard equations which assume the energy of system is constant. If we don't consider the effect of friction, the energy of system will not decrease and it becomes instable. So we must add a force in opposite direction of movement. Since the direction of movement is determined by v (velocity), the friction force is –bv. We later show that b also determines the bandwidth of each oscillator.
We model the state of each oscillator with the pair [x v], where x is the displacement and v is the velocity of particle. To write down the state variable equations of the system, we must compute the net force imposed to particle at each time. Knowing the mass of particle, say m, we can compute the acceleration of particle at each time, say a. so the state variable equation is simply:

$$\begin{bmatrix} x_{new} \\ v_{new} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t & 0 \\ 0 & 1 & \Delta t \end{bmatrix} \begin{bmatrix} x_{old} \\ v_{old} \\ a \end{bmatrix} \tag{1}$$

Where Δt is the inverse of sampling frequency.
The particle is imposed by three forces:

1. The diapason itself pulls the particle by force –kx.
2. The sound imposes a foreign force, say Fexternal.
3. The friction opposes to the movement by force –bv.

Now let us determine the value of each parameter. We know that to have a diapason with natural angular frequency w0, we must set $k = m\omega_0^2$ [6]. Viewing each diapason as a filter, one can ask the bandwidth of filter. [7] shows that the relation between bandwidth, say Δf and b is:

$$b = \frac{m\omega_0}{Q} \tag{2}$$

To compute Fexternal from the current sample we use the value of sample itself as the external force. Now we can compute a, using the following formula:

$$a = \frac{F - bv_{pr} - kx_{pr}}{m} \tag{3}$$

Up to now, we have enough information to compute the next state of each oscillator. As stated previously, our method does not need framing and energy of each diapason at each time can be calculated by the following formula:

$$E = \frac{1}{2}mv^2 + \frac{1}{2}kx^2 \tag{4}$$

For using this model in feature extraction after calculation of the energy for each of these oscillators, we use them as feature vectors in ASR systems.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Before using our model for feature extraction in a ASR system, we transform a speech with our human based model and compare it to spectrum domain of this speech. As shown in Figure 4, these two transformations have little differences. This comparing shows that this human based model can be used impressively in ASR systems. In addition, this method can be used as an effective and quick signal transformation instead of FFT or wavelet in various tasks.

4.1. Experiment conditions

The feature extraction algorithm proposed for speech recognition were tested on a English digit database. For training of the system we use 1386 digit sequences spoken by 18 speakers. In testing phase we use 200 digit sequences that uttered by speakers out of training database. The testing database split to four groups of 50 sequences and four types of noises added to these groups. These noises are car noise, exhibition noise, speech babble noise, and subway noise.
Recognition is performed using HTK[8] with a 16 emitting states and three mixture continuous HMM model (with 3-state silence model and single state inter-digit pause model). In the reference experiments, MFCC_0_D_A is used that consists of 13 standard cepstral coefficients including C0 augmented with first and second derivations of them. MFCC features were generated by applying a Hamming window of size 25 ms and overlap 10 ms to the same pre-emphasized 23-channel Mel-scale filterbank. The cepstral features were obtained from DCT of log-energy over the 23 frequency channels.

4.2. Experiment results

A complete set of experiments has been carried out for two different techniques: HEFE and MFCC. Table 1 summarizes the word error rates obtained based on the two features under different noises and under different SNR levels. Word error rates are obtained with an identically trained HMM system for

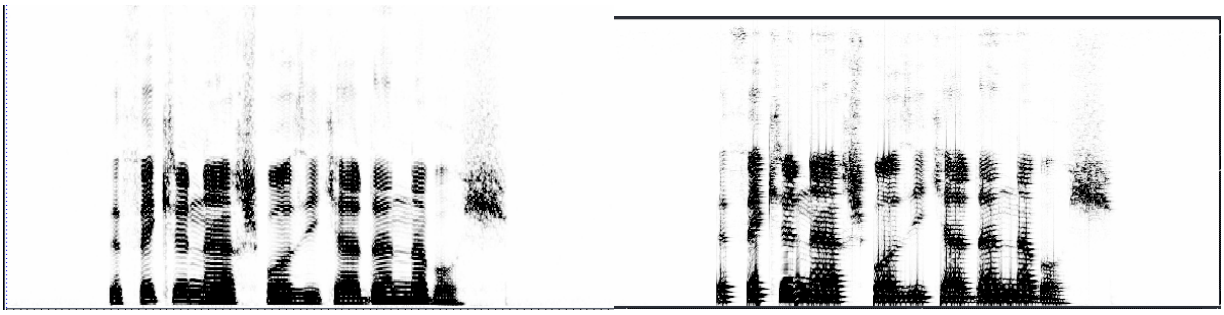

Figure 4*: Left Fig. demonstrate spectrum domain of a speech and right Fig. is transformation of this speech with our human ear based model.*

all models, and with speech subjected to additive car, exhibition, babble, and subway noise, at various SNR levels.

Table 1: *Comparing of MFCC and HEFE.*

Word Error Rate								
Noise	Feature	20 db	15 db	10 db	5 db	0 db	-5 db	Overall
Car	MFCC	13.29	24.05	53.16	80.38	93.67	93.67	59.7
	HEFE	13.29	23.42	55.06	74.05	82.28	93.67	56.96
Exhibition	MFCC	37.75	48.34	59.6	77.48	90.73	91.39	67.55
	HEFE	33.77	36.42	64.24	78.15	86.09	89.4	64.68
Babble	MFCC	26.95	44.31	72.46	88.02	94.61	99.4	70.96
	HEFE	23.95	34.73	63.47	76.05	83.23	88.62	61.68
Subway	MFCC	46.75	60.36	71.01	88.17	89.94	90.53	74.46
	HEFE	47.93	57.4	72.78	83.43	87.57	91.12	73.37

The following results can be inferred from the tabulated results:

1) For all contaminated speech, the human base ear models show superior performance for all noise types at most SNR levels.
2) For babble noise, HEFE demonstrate significantly better performance than MFCC.
3) For subway noise, improvements by the HEFE are least significant, but still noticeable.

5. . CONCLUSION

In this paper we have introduced a simple model for basilar membrane and hair calls based on physiological basis. We use this model for feature extraction in ASR systems. Simulations of the model using HMM recognizers trained on the digit database demonstrated the robustness of these human ear based features to different noises, significantly outperforming MFCC features at babble noise.

6. REFERENCES

[1] Rabiner L.R. and Shafer R.W. ,*Digital Processing of Speech Signals,* Englewood Cliffs, New Jersey: Prentice-Hall, 1987.

[2] Davis S.B. and Mermelstein P., "Comparison of Parametric Representation of Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. ASSP* 28(4): 357-366, 1980.

[3] Furui S. "Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", *IEEE Trans. ASSP* 34(1): 52-59, 1986.

[4] Kronland-Martinet R. and Grossmann A., "Application of time-frequency and time-scale methods (wavelet transform) to the analysis, synthesis and transformation of natural sounds", *Representations of musical signals, MITPress*: 45-85, Cambridge,1991.

[5] Douglas O'Shaughnessy, *Speech Communication, Human and Machine*, pp. 128-163, Addision-Wesley Publishing Company, U.S.A, 1987.

[6] C. H. Edwards, D. E. Penney, *Differential Equations and Boundary Value Problems: Computing and Modeling*, 2nd Edition, Prentice Hall, 1999.

[7] H. D. Young, R. A. Freedman, *University Physics with Modern Physics with Mastering Physics*, 11th Edition, Benjamin Cummings, 2004.

[8] S. Young, G. Everman, D. Kershaw, G Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book (for HTK Version 3.2)*, Cambridge University, 2002.