The Place Theory as an Alternative Solution in Automatic Speech Recognition Tasks

José Luis Oropeza-Rodríguez[1], Sergio Suárez-Guerra [1],
and Mario Jiménez-Hernández[2]

[1] Computing Research Center, National Polytechnic Institute,
Juan de Dios Batiz s/n, P.O. 07038, Mexico
[2] ESIME Zacatenco, National Polytechnic Institute,
Av. Politécnico, P.O. 07038, Mexico
{joropeza,ssuarez}@cic.ipn.mx, mjimenezh@ipn.mx

Abstract. Recently the parametric representation using cochlea behavior has been used in different studies related with Automatic Speech Recognition (ASR). This paper shows how using an alternative solution reported in the state of the art solves the Lesser and Berkeley's cochlea model in ASR tasks. An approach that considers a new form to construct the bank filter in the parametric representation used to extract MFCC is proposed. Then this distribution of the bank filter to have a new representation of the speech in frequency domain is used. It is important to indicate that MFCC parameters use Mel scale to create a bank filter. The cochlea behavior based on the theory to create the central frequencies of the bank filter was used, .The Mel scale function was substituted for our purpose. A 98.5% performance was reached, for a task that uses isolated digits pronounced by 5 different speakers in the Spanish language and corpus SUSAS with neutral sound records with some advantages in comparison with MFCC was used.

Keywords: Automatic Speech Recognition, Speech recognition, cochlea operation, place theory and bank filter component.

1 Introduction

For a long time Automatic Speech Recognition Systems have used parameters related with Cepstrum and Homomorphic Analysis of Speech [1], Linear Prediction Coefficient (LPC) [2], Mel Frequency Cepstrum Coefficients (MFCC) [3], Perceptual Linear Prediction (PLP)[4]. One important aspect to mention is that cochlea properties have not been considered in the models mentioned above. However, recently works related with the application of the cochlea behavior in ASR systems can be found because in recent years researchers have been emphasizing "human engineering", that is, to adopt the processing strategies of the human auditory perception. The application of such human perceptual feature may improve ASR performance which has been established in literature [5][6][7] [8][9][10]. In [10] an extraordinarily precise auditory model was used to extract the excitation dependent shapes of the delay trajectories and then t a set of features without any other spectral information were used to carry out speech recognition task under different noise conditions on the

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 167–174, 2014.

TIMIT database. However, average recognition rates do not reach that of the MFCC features (except for very low noise SNRs), but the system behaves very stable under different noise conditions. In [9] they proposed a feature extraction method for ASR based on the differential processing strategy of the AVCN, PVCN and the DCN of the nucleus cochlear. The method utilized a zero-crossing with peak amplitudes (ZCPA) auditory model as a synchrony detector to discriminate the low frequency formants. They used HMM recognition of isolated digits that showed improved recognition rates in clean and in non- stationary noise conditions than the existing auditory model. In [8] they employed a counterpart of the next physiological processing step in comparison with frequency decomposition and compression of amplitudes concepts. A simplified model of short-term adaptation into MFCC feature extraction was incorporated. The proposal mentioned above was compared with the structurally related RASTA, CMS and Wiener filtering and performs well in combination with Wiener filtering. Compared to the structurally related RASTA, the adaptation model provides superior performance on AURORA 2, and, if Wiener filtering is used prior to both approaches, on AURORA 3 as well.

On the other hand, the most important organ in human hearing is the cochlea and various phenomenological (as used by Ghitza where basilar membrane is modeled by a gammatone bank filter) and physiological models have been proposed for a long time [11][12]. At same time MFCC have been used for different tasks of ASR and speech representation. This paper uses a physiological model proposed in literature instead of phenomenological models that have been used for same application. An important difference is that phenomenological models are based on bank filters, as EIH (Ensemble Interval Histogram Model was conceived by Ghitza) while physiological models use fluid mechanics [13];

2 Characteristics and Generalities

The cochlea is a long, narrow, fluid-filled tunnel which spirals through the temporal bone. This tunnel is divided along its length by a cochlear partition into an upper compartment called scala vestibuli (SV) and lower compartment called scala timpani (ST). At the apex of the cochlea, SV and ST are connected to each other by the helicotrema [14]. A set of models to represent the operation of the cochlea has been proposed [15][16][17][18]. In mammals, vibrations of the stapes set up a wave with a particular shape on the basilar membrane. The amplitude envelope of the wave is first increasing and then decreasing, and the position at the peak of the envelope is dependent on the frequency of the stimulus [19]. The amplitude of the envelope is a two-dimensional function of distance from the stapes and frequency of stimulation The curve shown in Fig. 1 is a cross-section of the function for fixed frequency. Frequency responses analyzed by Von Békésy are shown in Fig. 1, where each part of the basilar membrane responds maximally to a certain frequency, and as the frequency increases so does the maximum place of the envelope. If low frequencies excite the cochlea, the envelope is nearest to the apex, but if high frequencies excite it, the envelope is nearest to the base.

This paper proposes an equation extracted from the fluid mechanical model to find a relationship between these frequencies and the place of the excitation into the cochlea. With that value a new distribution of the bank filter to extract parameters for ASR tasks is proposed.

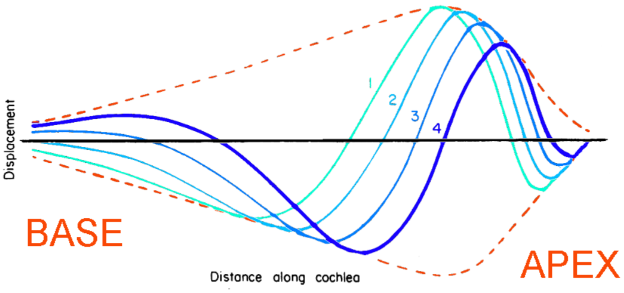

Fig. 1. Wave displacement inside cochlea

Let $u = (u_1, u_2, u_3)$ be the fluid velocity, p the pressure, and ρ the constant density of the fluid. The mass of fluid in a fixed volume V can change only in response to fluid flux across the boundary of the volume. Thus [23],

$$\frac{d}{dt}\int_V \rho dV = -\int_S \rho(u \bullet n)dS = 0 \tag{1}$$

Where S is the surface of V, and $n = (n_1, n_2, n_3)$ is the outward unit normal to V.

After considering that the momentum of the fluid in a fixed domain V can change only in response to applied forces or to the momentum flux across the domain boundary, and using the divergence theorem to convert surface integrals to volume integrals, 2 is obtained:

$$\int_V \left(\rho \frac{\partial u_i}{\partial t} + \rho \nabla \bullet (u_i u) + \frac{\partial p}{\partial x_i} \right) dV = 0 \tag{2}$$

After considering that V is arbitrary, fluid motions are of small amplitude and there is an irrotational flow, the following equations are shown:

$$\begin{gathered} \rho \frac{\partial \phi}{\partial t} + p = 0, \\ \nabla^2 \phi = 0 \end{gathered} \tag{3}$$

Lesser and Berkley developed a model that combines these last two equations with the equation of a damped, forced harmonic oscillator and is considered one of the simplest of the cochlea models. They propose that each point of the basilar membrane is modeled as a simple damped harmonic oscillator with mass, damping, and stiffness that vary along the length of the membrane. Thus, the movement of any part of the membrane is assumed to be independent of the movement of neighboring parts of the membrane, as there is no direct lateral coupling. The deflection of the basilar membrane, η(x, t), is specified by a model of a forced harmonic oscillator defined as

$$m(x)\frac{\partial^2 \eta}{\partial t^2} + r(x)\frac{\partial \eta}{\partial t} + k(x)\eta = p_2(x, \eta(x,t), t) - p_1(x, \eta(x,t), t) \tag{4}$$

Where $m(x) = 0.1,\ \ r(x) = 300e^{-ax},\ \ k(x) = 10^9 e^{-2ax}$. An analytical solution of this problem can be found using standard Fourier series [23]. Solutions of this form are looked for:

$$\phi = x\left(1 - \frac{x}{2}\right) - \sigma y\left(1 - \frac{y}{2\sigma}\right) + \sum_{n=0}^{\infty} A_n \cosh[n\pi(\sigma - y)]\cos(n\pi x) \tag{5}$$

3 Auditory Model

This paper proposes solving the Lesser and Berckley equation using the solution proposed in [20]. This solution is related with the place theory of hearing, initially proposed by Von Békésy. To perform the analysis each section of the membrane is considered as a forced harmonic isolated oscillator , which is excited by an external force $Fe^{j\omega t}$ that represents the driving force on each section of the basilar membrane and this force is produced by vibrations transmitted into the cochlea by the oval window. Two solutions are proposed related with the before mentioned equation. Firstly, the forced harmonic oscillator is represented by the following equation

$$m(x)\frac{d^2\eta}{dt^2}+R_m(x)\frac{d\eta}{dt}+k(x)\eta=Fe^{j\omega t} \tag{6}$$

Where m is the mass, R_m mechanical resistance and k is the damping constant. Considering that $\eta = Ae^{j\omega t}$, then amplitude of the wave sound into the cochlea is represented by [20]. Secondly, a damped harmonic oscillator with the following equation is considered:

$$m(x)\frac{d^2\eta}{dt^2}+R_m(x)\frac{d\eta}{dt}+k(x)\eta=0 \tag{7}$$

Then, a solution is given by

$$\eta = Ae^{-\beta t}\cos(\omega_0 t+\phi) \tag{8}$$

Equation 9 shows that the amplitude for each section of the membrane depends of the frequency ω in the applied force. The amplitude has a maximum when the denominator has its minimum value and this occurs at a specific frequency excitation called resonance frequency. This is defined by the values of mass and stiffness, when the frequency ω of the applied force is equal to $k(x)/m(x)$ it is said that the system is resonant in amplitude and obtains the maximum value of the basilar membrane displacement. This last equation can be expressed as a function of frequency and distance, if considering that $\omega = 2\pi f$ thus, this is possible using our purpose Literature does not find an equal relationship [20].

$$A=\frac{F/m(x)}{\sqrt{\left(4\pi^2f^2-\frac{k(x)}{m(x)}\right)^2+4\pi^2f^2\frac{R_m(x)^2}{m(x)^2}}} \tag{9}$$

4 Experiments and Results

From the last equation a computational model was developed to obtain the distance where the maximum displacement of the basilar membrane to a specific excitation frequency of the system occurs, which depends of the physical characteristics of the basilar membrane. The following procedure describes the computational model of the cochlea using this propose [21]. It’s important to mention that the maximum response of the pressure curve used in [20]was obtained. Firstly, 5 speakers pronounced 10 digits from 0 to 9; Spanish digits were used as a workbench that is "cero, uno, dos, tres, cuatro, cinco, seis, siete, ocho and nueve". LPC, MFCC, CLPC were used and our coefficients named EPCC (Earing Perception Cepstrum Coefficients), obtaining better percent correct recognition in some tasks using them in comparison with others representations mentioned above. HTK Hidden Markov Model Toolkit were used as training and recognition software; our new parameters were added into HSigp.c file, contained inside HTK http://htk.eng.cam.ac.uk, and used in task of ASR employing HTK.

PROCEDURE DESCRIPTION

- Obtain speech signal, realize preprocessing (It includes pre-emphasis, segmentation, windowing and feature extraction), for each sentence.
- The feature extraction, used the same procedure as MFCC but filter bank is constructed following the next steps.
  - Take the minimal and maximal frequency where filter bank are going to be constructed.
  - Calculate maximal and minimal distance from the stapes of the cochlea, nearer to start implies high frequencies, farthest implies low frequencies.
  - Determine a set of distances equally spaced
  - Determine the frequency related with these distances, this represents the center of the filter bank.
  - Construct filter bank with frequency center obtained from the analysis of the Neely model using values in table 1.
- Follow the same steps to obtain MFCC, multiply spectral representation from Fourier Transform with filter bank, calculate energy by bands using logarithm, and finally, apply discrete cosine transform.
- Obtain a new set of coefficients for each speech signal.
- Train the ASR and proceed with recognition task using the new parameters.

This first experiment used a database that contains only digits in the Spanish language. The characteristics of the samples were frequency sample 11025, 8 bits per sample, PCM coding, mono-estereo. The evaluation of the experiment proposed involved 5 people (3 men and 2 women) with 300 speech sentences to recognize for each one (100 for training task and 200 for recognition task were used). About 1500 speech sentences extracted from 5 speakers individually were taken, and the Automatic Speech Recognition using Hidden Markov Models was trained with 4 (2 states with information and 2 dummies to connection with another chain), 5 (3 states with information and 2 dummies to connection with another chain) and 6 states (4 states with information and 2 dummies to connection with another chain). Also, 3 Gaussian Mixture for each state in the chain Markov were employed. The parameters extracted of the speech signal were 39 (13 MFCC, 13 delta and 13 energy coefficients) When using MFCC or our proposal, they are used to train the Hidden Markov Model. Table 1 contains results obtained in percentage when using LPC, CLPC, MFCC and our parametric representation as parameters to training. Table 2 shows results using also Delta and Acceleration coefficients. It is important to remember that HTK give us results in two forms: by sentence and by words http://htk.eng.cam.ac.uk. Table 3 contains results obtained in percentage when using LPC, CLPC and MFCC, DELTA, ACCELERATION AND THIRD DIFFERENTIAL. The headings in table 1, 2 and 3 represent the number of states used for each HMM used in the experiments.

Table 1. LPC, CLPC and MFCC coefficients

# STATES	*4*	*5*	*6*
LPC SENTENCE	87.5	94	94
CLPC SENTENCE	90	97.5	98.5
MFCC SENTENCE	97.5	97	99
OUR PROPOSAL	99.25	99.35	99.6
LPC WORDS	87.94	94.47	94.47
CLPC WORDS	90.45	97.99	98.99
MFCC WORDS	97.99	97.49	99.5
OUR PROPOSAL	99.35	99.45	99.75

Table 2. LPC, CLPC, MFCC, DELTA AND ACCELERATION coefficients

# STATES	*4*	*5*	*6*
LPC SENTENCE	79	90.5	91.5
CLPC SENTENCE	93	99	99
MFCC SENTENCE	99	99	99
OUR PROPOSAL	99.30	99.6	99.7
LPC WORDS	79.4	99.4	91.96
CLPC WORDS	93.47	99.5	99.5
MFCC WORDS	99.5	99.5	99.5
OUR PROPOSAL	99.45	99.75	99.8

Table 3. LPC, CLPC, MFCC AND DELTA, ACCELERATION, DELTA, AND THIRD DIFFERENTIAL coefficients

# STATES	*4*	*5*	*6*	# STATES	*4*	*5*	*6*
LPC SENTENCE	77	89.5	89	LPC WORDS	77.39	89.95	89.45
CLPC SENTENCE	89.5	99	99	CLPC WORDS	89.95	99.5	99.5
MFCC SENTENCE	98.5	99	99	MFCC WORDS	98.99	99.5	99.5
OUR PROPOSAL	99.4	99.6	99.8	OUR PROPOSAL	99.6	99.8	99.8

Table 4. Results obtained using HTK, Susas Corpus and manual labeling

	MFCC		*EPCC*	
	sen-tence	*word*	*sen-tence*	*word*
boston1	91.84	92.06	90.2	90.84
boston2	95.51	95.63	93.88	94.05
boston3	96.73	96.83	92.65	92.86
general1	96.73	96.83	95.51	95.24
general2	94.29	94.44	93.06	93.25
general3	93.47	93.65	94.69	94.84
nyc1	91.84	92.06	93.06	92.86
nyc2	91.02	91.27	89.8	90.08
nyc3	95.92	96.03	90.2	90.48

Table 5. Results obtained using HTK, Susas Corpus and automatic labeling based in zero crossing and energy

	MFCC		*EPCC*	
	sen-tence	*word*	*sen-tence*	*word*
boston1	93.47	93.47	91.43	91.43
boston2	97.55	97.55	96.33	96.33
boston3	99.18	99.18	97.14	97.14
general1	95.92	95.92	95.1	95.1
general2	95.92	95.92	92.24	92.24
general3	91.84	91.84	89.8	89.8
nyc1	93.88	93.88	95.51	95.51
nyc2	98.37	98.37	97.96	97.96
nyc3	97.14	97.14	90.2	90.48

Secondly, a corpus elaborated by J. Hansen at the University of Colorado Boulder was used. He has constructed database SUSAS (Speech Under Simulated and Actual Stress) http://catalog.ldc.upenn.edu/LDC99S78. Only 9 speakers, with ages ranging from 22 to 76, named “boston1”, “boston2”, “boston3”, “general1”, “general2”, “general3”, “nyc1”, “nyc2” and “nyc3”were used. Normal corpus not under Stress sentences contained into corpus were applied. The words were “brake, change, degree, destination, east, eight, eighty, enter, fifty, fix, freeze, gain, go, hello, help, histogram, hot, mark, nav, no, oh, on, out, point, six, south, stand, steer, strafe, ten, thirty, three, white, wide, & zero”. A total of 4410 files of speech were processed. Finally, Table 4 and 5 show results when using our proposal (Earing Perceptual Cepstrum Coefficients –EPCC-) the best representations used in the state of the art and in the last experiment versus MFCC in Susas corpus. The heading of each line in tables 4 and 5 represent each speaker of the Susas Corpus used in this experiment.

5 Conclusions and Future Works

As shown in this paper a new parameter for ASRs task has been described. They employ the functionality of the most important organ for humans and mammalians in hearing, the cochlea. At this moment all investigations are oriented to a set of models that use pronounced speech signals or frequency domain behavior, considering perceptual effects in humans. However, they do not consider the function principle of the hearing phenomena that occurs in the inner ear. This proposal with the results obtained has been integrated into the ASRs task satisfactorily to reach a performance of 99.8% in digit task for the Spanish language and 93.36 in Susas Corpus. This demonstrates the cochlea functionality for extracting information from the speech signal. It can be compared with another database such as TIMIT to test the robustness of the results.

Acknowledgment. We want to acknowledge The National Polytechnic Institute (IPN, Mexico), also Center for Computing Research and especially SIP project numbers 20141310 and 20141454 for their support.

References

1. Noll, A.M.: Shortime Spectrum and Cepstrum Techniques for Vocal Pitch Detection. Journal of Acoustical Society of America 36, 296–302 (1964)
2. John, M.: Linear Prediction: A Tutorial Review. Proceedings of the IEEE 63(4), 561–580 (1975)
3. Davis, S.B., Mermelstein, P.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentence. IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-28(4) (1980)
4. Hermansky: Perceptual Linear Predictive (PLP) analysis of speech. Journal of Acoustical Society of America, 1738–1752 (April 1990)
5. Kim, D.S., Lee, S.Y., Kill, R.M.: Auditory processing of speech signals for robust speech recognition in real word noisy environments. IEEE Trans. Speech Audio Processing 7(1), 55–69 (1999)

6. Geisler, C.D.: A model of the effect of outer hair cell motility on cochlear vibration. Hear. Res. 24, 125–131 (1996)
7. Geisler, C.D., Shan, X.: A model for cochlear vibration based on feedback from motile outer hair cells. In: Dallos, P., Geilser, C.D., Matthews, J.W., Ruggero, M.A., Steele, C.R. (eds.) The Mechanics and Biophysics of Hearing, pp. 86–95. Springer, New York (1990)
8. Holmberg, M., Gelbart, D., Hemmert, W.: Automatic speech recognition with an adaptation model motivated by auditory processing. IEEE Trans, Audio, Speech, Language Processing 14(1), 44–49 (2006)
9. Haque, S., Togneri, R.: A feature extraction method for automatic speech recognition based on the cochlear nucleus. In: 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30 (2010)
10. Harczos, T., Szepannek, G., Klefenz, F.: Towards Automatic Speech Recognition based on Cochlear Traveling Wave Delay Trajectories, Auditory signal processing in hearing-impaired listeners. In: Dau, T., Buchholz, J.M., Harte, J.M., Christiansen, T.U. (eds.) 1st International Symposium on Auditory and Audiological Research, ISAAR 2007 (2007) ISBN: 87-990013-1-4. Print: Centertryk A/S
11. de Boer, S.E.: Mechanics of the cochlea: modeling effects. In: Dallos, P., Fay, R.R. (eds.) The Cochlea, ch. 5. Springer, USA (1996)
12. Robles, L., Ruggero, M.A.: Mechanics of the Mammalian Cochlea. Physiological Reviews 81(3) (July 2001), Printed in USA
13. Peterson, Bogert: A dynamical theory of the cochlea. Journal of the Acoustical Society of America 22(3), 369–381 (1950)
14. Keener, Sneyd: Journal of Mathematical Physiology. Springer, USA (2008)
15. Elliot, S.J., Ku, E.M., Lineton, B.A.: A state space model for cochlear mechanics. Journal of Acoustical Society of America 122, 2759–2771 (2007)
16. Elliott, S.J., Lineton, B., Ni, G.: Fluid coupling in a discrete model of cochlear mechanics. Journal of Acoustical Society of America 130, 1441–1451 (2011)
17. Ku, E.M., Elliot, S.J., Lineton, B.A.: Statistics of instabilities in a state space model of the human cochlea. Journal of Acoustical Society of America 124, 1068–1079 (2008)
18. Neely, S.T.: A model for active elements in cochlear biomechanics. Journal of Acoustical Society of America 79, 1472–1480 (1986)
19. Békésy: Concerning the pleasures of observing and the mechanics of the inner ear, Nobel Lecture (December 11, 1961)
20. Mario, J.H., Rodríguez, J.L.O., Guerra, S.S., Barrón, R.: Fernández: Computational Model of the Cochlea using Resonance Analysis. Journal Revista Mexicana Ingeniería Biomédica 33(2), 77–86 (2012)
21. Hernández, J.: Mario: Modelo mecánico acústico del oído interno en reconocimiento de voz, Ph. D. Thesis, Center for Computing Research-IPN (June 2013)
22. Lesser, M.B., Berkley, D.A.: Fluid mechanics of the cochlea. Journal Fluid Mechanics 51(Pt. 3), 497–512 (1972)