New parameters for Automatic Speech Recognition based on the mammalian cochlea Model using resonance analysis

José Luis Oropeza Rodríguez

Department of Digital Signal Processing, Center for Computing Research, National Polytechnic Institute, Mexico, DF. Instituto Politécnico Nacional, Centro de Investigación en Computación CIC.
joropeza@cic.ipn.mx

Abstract

Undoubtedly the compact representation by a set of Mel Frequency Cepstrum Coefficients (MFCC) has been used satisfactorily for ASR [9]. The cochlea is an organ, in humans or mammals that converts the frequency perceived by the ear in punctual stimulation to excite the nerve auditory that receives a set of stimulus that comes from speech sound pressure. A new approach is proposed that considers this phenomenon to construct the bank filter in our parametric representation. Then we substitute the distribution of the bank filter in the Mel scale function for a different distribution that depends of the inner ear response to the stimulus that it receives. The place theory is used which achieves a 99.8% performance.

Finally, this paper compares the performance of different acoustic representations in Continuous Automatic Speech Recognition system (CASRs) based on words. The cochlea operation is explained that permits obtaining a model and we will show that one alternative solution to the model based on fluid mechanical proposed by Lesser and Berkley [19], can be obtained if resonance analysis weather Fourier series is used as a solution.

Index Terms: Speech recognition, cochlea operation, place theory and bank filter.

1. Introduction

For a long time Automatic Speech Recognition Systems have used parameters related with Cepstrum and Homomorphic Analysis of Speech [27], [28], [29] and [30] Linear Prediction Coefficients (LPCs) [22], Mel Frequency Cepstrum Coefficients (MFCCs) [38], Perceptual Linear Prediction Coefficients (PLPs) [14], these last two being the most important. This set of parameters uses spectral representation as the most significant representation of the speech signal. In each of these representations, the principal objective is to have a representation to compress the speech data without irrelevant information pertinent to the phonetic data analysis and to enhance aspects of the signal that contribute significantly to the detection of phonetic differences. Other tasks where the reduction of the information of the speech signal is relevant are there when a great amount of reference information, such as speech signals for ASR that employed digital networks, is stored. Then, the reduction in the capacity of this information is a problem when we process database speech, used for transmission or storage [35] and [38].

Undoubtedly, speech is the most important auditory information perceived by humans, but the auditory system does not respond as a linear but a logarithmic measurement system. MFCC and PLP coefficients employ Mel and Bark scales respectively that consider perceptual aspects to obtain a set of coefficients that represent the speech signal. One aspect to mention is that cochlea properties have not been considered. Inside the cochlea a particular frequency analysis is realized. It transforms frequency response into distance response [32]. Then, the solutions before mentioned take only the perceptual response without considering the principal operation of the cochlea. Therefore to understand the cochlear operation using models permits an analysis of speech signal closer to hearing human.

On the other hand, the most important organ in human hearing is the cochlea and various phenomenological and physiological models have been proposed for a long time. [5], [6], [7], [8] and [37]. At same time MFCC have been used for different tasks of ASR and speech representation; the perceptual effect is interesting because from the psychological and physiological aspect there is an adequate form of operation of the speech signal behavior when the human perceive it. For another side, since 1930’s and after 1950’s, the analysis and study of the cochlea behavior has generated many publications and as a consequence the implementation of the devices related with the form that the human can hear the sounds (cochlear implants) [16], [20], [21], [23], [24] and [31]; and considered aspects related with the audition into inner ear that has a capability to divide the sound coming to the outer and medium ear and process the sound that it captures to divide into a set of frequencies that interact with ciliós into the basilar membrane which is an important part of inner ear functionality.

In these studies a set of models to represent the operation of the cochlea has been proposed [1], [10], [11], [13], [18], [25], [26], [33], [36], [40], and [41]. This paper proposes new parameters, that were used for ASR tasks, that are related with the cochlea model proposed by Lesser and Berkeley [19] based on the fluid mechanical. They are based on an alternative solution of the last model mentioned above, used in resonance analysis.

A database that contains spontaneous words for applications based in commands was used [13]. The sentences were pronounced by various speakers to probe the reliability of the parameters. Although variation among speakers is an important problem in its right, attention is focused here on speaker-dependent representations to restrict the different sources of variation in the acoustic data. For a long time, the bandpass filtering and linear prediction coding approaches have been used satisfactory with pattern recognition techniques such as Dynamic Programming Time Alignment, Dynamic Time Warping, Hidden Markov Models, Neural Networks; among others [9].

In this experiment 5 speakers pronounced 10 phonetically similar words; Spanish digits were used as a workbench [34]. LPC, MFCC, CLPC and coefficients were used, and obtained better percent correct recognition in some tasks using them in...
comparison with other representations mentioned above. HTK Hidden Markov Model Toolkit as training and recognition software was used and these new parameters were added into a HSigp.c file and used in the task of ASR employing HTK [42].

The solution based on resonance analysis permits having an equation that relates direct frequency with distance where cochlea is stimulated [15]. The Lesser and Berkley model was selected because it is based on fluid mechanics and that exactly occurs in the cochlea. Another analysis such as Neely’s model can be used [15] and [26], that is based on the macro-mechanics and micromechanics solution for an analysis of the cochlea behavior. However, it is difficult to find an equation that can obtain directly a relation between frequency and distance with Neely’s model.

2. Theory aspects

The cochlea consists of a spirally coiled tube divided longitudinally into three sections, one being the cochlear duct and is relatively small in a cross-sectional area named scala media, and the other two, the scala vestibula, and scala tympani are larger and roughly equal in area. These structures are communicating, fluid-filled spaces within the temporal bone. The cochlea is the primary receptor organ of hearing; it receives acoustic signals from the middle-ear ossicles and distributes sound information to individual auditory nerve fibers. The separation between scala media and scala vestibuli (Reissner’s membrane that is very thin) provides a chemical barrier, but is probably not important to cochlear mechanics.

The separation between scala media and scala tympani (organ of Corti) contains the final receptor cells (inner hair cells) and plays an important role in cochlear mechanics. Inside the Organ of Corti, there is the Basilar membrane that is much thicker and supports it. As far as this work is concerned, the cochlear duct is dynamically part of the basilar membrane, and experimental evidence of Békésy shows that the Reissner membrane vibrates in time with the basilar membrane [2] and [3]. The cochlear partition is mechanically tuned such that higher frequency tones cause localized vibrations nearer to the stapes. The stiffness of the partition comes primarily from the basilar membrane which spans the gap between scala media and scala tympani. The sensory hair cells are situated on the basilar membrane and move as the basilar membrane moves. Over the top of the hair cells lies the tectorial membrane. When the cochlear partition vibrates, a shear motion between the basilar and tectorial membranes will bend the tufts of hair (stereocilia) at the top of the hair cells. The bending of the stereocilia modulates intracellular potential, thereby accomplishing mechanical-to-electrical transduction [24].

The ear as a sensory organ is far more complex than other sensory organs. The sensory cells are located in the cochlea but the cochlea not only serves to convert sound into a code of neural impulses in the auditory nerve, but also performs the first analysis of sounds that prepares them for further analysis in the auditory nervous system. This analysis consists primarily of separating sounds into bands of frequencies before they are coded in the discharge pattern of individual auditory nerve fibers. The separation of sounds is accomplished by the properties of the basilar membrane and the sensory cells that are located along its length. The cochlea is more frequency selective for weaker rather than louder sounds, which facilitates the detection of weak sounds [39].

The cochlea also compresses the amplitudes of sounds, which makes it possible to code sounds within the very large range of sound intensities that is covered by normal hearing. Without such amplitude compression the ear could not detect and analyze sounds in the intensity range of normal hearing [24].

In mammals, stapes vibrations set up a wave with a particular shape on the basilar membrane. The amplitude envelope of the wave first increases, then decreases, and the position of the peak of the envelope, depending on the frequency of the stimulus (von Békésy, 1960) [2] and [3]. The wave speed decreases as it moves along the membrane, resulting in a continual decrease in time, and an apparent increase in frequency. Low-frequency stimuli have a wave envelope that peaks closer to the apex of the cochlea (i.e., near the helicotrema), and as the frequency of the stimulus increases, the envelope peak moves toward the cochlea base.

The amplitude of the envelope is a two-dimensional function of distance from the stapes and stimulation frequency. Another way to present the data is to give cross-sections for a fixed distance. This gives the envelope amplitude a function of frequency, for a fixed distance from the stapes, i.e., the frequency response of the basilar membrane for that fixed distance. Frequency responses were measured by Von Békésy, from which it can be seen that each part of the basilar membrane responds to a certain frequency, and as the frequency increases, the site of maximum response moves toward the stapes. In this way the cochlea determines the frequency of the incoming signal from the place on the basilar membrane of maximum amplitude, the so-called place theory of hearing.

Let \( \mathbf{u} = (u_1, u_2, u_3) \) be the fluid velocity, \( p \) the pressure, and \( \rho \) 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 [17],

\[
\frac{d}{dt} \int \rho dV = -\int \rho \mathbf{u} \cdot \mathbf{n} dS = 0
\]

(1)

Where \( S \) is the surface of \( V \), and \( \mathbf{n} = (n_1, n_2, n_3) \) 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:

\[
\int \left( \rho \frac{\partial u_i}{\partial t} + \rho \mathbf{u} \cdot \nabla u_i - \frac{\partial p}{\partial x_i} \right) dV = 0
\]

(2)

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

\[
\rho \frac{\partial \phi}{\partial t} + p = 0,
\]

(3)

\[
\nabla^2 \phi = 0
\]

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 [19]. They propose that each point of the basilar membrane is modeled as a simple damped harmonic oscillator.

1733
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, \( \eta(x, t) \), is specified by a model of a forced harmonic oscillator defined as

\[
m(x) \frac{\partial^2 \eta}{\partial t^2} + R_m(x) \frac{\partial \eta}{\partial t} + k(x) \eta = F(x, t) \tag{4}
\]

An analytical solution of this problem can be found using standard Fourier series [19]. Solutions of this form are looked for:

\[
\phi = \sum_{n=1}^{\infty} A_n \cos(n\pi(x - y)) \cos(n\pi t) \tag{5}
\]

This paper proposes solving the Lesser and Berckley equation using Resonance analysis [15]. This solution is related with the place theory of hearing, initially proposed by Von Békésy [2] and [3]. To perform the analysis each section of the membrane is considered as a forced harmonic oscillator isolated, which is excited by an external force \( F_{ext} \) 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

\[
m(x) \frac{d^2\eta}{dt^2} + R_m(x) \frac{d\eta}{dt} + k(x)\eta = F_0 e^{i\omega t} \tag{6}
\]

Where \( m \) is the mass, \( R_m \) mechanical resistance and \( k \) is the damping constant. Considering that \( \eta = A_0 e^{i\omega t} \), then amplitude of the wave sound into the cochlea is represented by [15]. As can be seen, it depends on a set of parameters related with mechanics impedance, such as used commonly by [33] and [26], among others. Also, it is related with the observation mentioned by Békésy, in the sense that a rubber membrane can be made whose lateral tension is such that a point pressed into the membrane produces an elliptical bulge; this membrane behaves in exactly the same way described by the resonance theory [2] and [3].

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

\[
\eta = A_0 e^{-\beta t} \cos(\omega_0 t + \phi) \tag{8}
\]

And therefore the displacement of each membrane section is defined by the real part of the last equation.

\[
\eta = \frac{F}{\omega_0 Z_m(x)} \sin(\omega_0 t - \Theta(x)) \tag{9}
\]

\( Z_m \) is the mechanical impedance of the system. In equation 9 the amplitude is defined by \( A = F / \omega_0 Z_m(x) \) and can be expressed algebraically in terms of mass, damping and stiffness, by the following expression

\[
A = \left| \frac{F/m(x)}{\left(4\pi f^2 - k(x)/m(x)\right)^2 + 4\pi^2 f^2 R_m(x)/m(x)} \right| \tag{10}
\]

Equation 11 shows that the amplitude for each section of the membrane depends of the frequency \( \omega_0 \) 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 \( \omega_0 \) 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 \( \omega_0 = 2\pi f \) [15] thus, this is possible using our propose, in literature is not find a relation equal

\[
A = \frac{F/m(x)}{\left(4\pi f^2 - k(x)/m(x)\right)^2 + 4\pi^2 f^2 R_m(x)/m(x)} \tag{11}
\]

### 3. Discussion

From this last equation a computational model was developed to obtain the distance where the maximum displacement of the basilar membrane occurs to a specific excitation frequency of the system, which depends on the physical characteristics of the basilar membrane. The following procedure describes the computational model of the cochlea using resonance analysis.

1. Obtain speech signal, perform pre-processing (It includes pre-emphasis, segmentation, windowing and feature extraction), for each sentence.
2. In the feature extraction, the same procedure as MFCC is used but a filter bank is constructed following these steps:
   2.1 Take the minimal and maximal frequency where the filter bank is going to be constructed.
   2.2 Calculate maximal and minimal distance from the stapes of the cochlea, nearer to start implies high frequencies, fartherst implies low frequencies.
   2.3 Determine a set of distances equally spaced between minimal and maximal distance. Also the difference between them.
   2.4 Calculate the frequency related with these distances as they represent the center of the filter bank. Applying 11. The set of parameters used in equation 11 are obtained from [33], where mass, mechanical resistance, stiffness and damping are mentioned. Then the substitution of these parameters leads to an equation without variables.
2.5 Construct filter bank from centers obtained in step 2.4

2.6 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.

3. Obtain a new set of coefficients for each speech signal.

4. Train the ASR and proceed with recognition task using the new parameters.

An experimental database that contains only digits in Spanish language was used [34]. The characteristics of the samples were frequency sample 11025, 8 bits per sample, PCM coding, mono-stereo. The evaluation of the experiment proposed involved 5 people (3 men and 2 women) with 300 speech signals were processed to recognize for each. Speech signals were recorded at in laboratory environment that signified that they were clean speech without noise. After that, speech signals were processed to eliminate information not useful that in the speech sentences were processed to find start and end points of speech signal, using software. This software used the energy parameter to find it the interesting region.

Firstly, 1500 speech sentences extracted from 5 speakers individually were used (100 for training task and 200 for recognition task), and trained the Automatic Speech Recognition using Hidden Markov Models with 6 states (4 states with information and 2 dummies to connect with another chain) [4]. Also, 3 Gaussian Mixture for each state in the Markov chain was employed. The parameters extracted from the speech signal were 39 (13 MFCC, 13 delta and 13 energy coefficients), they are used to training the Hidden Markov Model. Table 1 contains percentage results obtained when using LPCs, CLPCs, MFCCs and our coefficients as parameter training. Table 2 shows results using also delta and acceleration coefficients. Table 3 contains results obtained in percentage when using LPCs, CLPCs and MFCCs and our propose using DELTA, ACCELERATION AND THIRD DIFFERENTIAL. Finally, in word recognition using phonemes 100% with MFCC and our algorithm were obtained.

<table>
<thead>
<tr>
<th>STATE NUMBER used in HMM/coefficients</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC SENTENCE</td>
<td>87.5</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>CLPC SENTENCE</td>
<td>90</td>
<td>97.5</td>
<td>98.5</td>
</tr>
<tr>
<td>MFCC SENTENCE</td>
<td>97.5</td>
<td>97</td>
<td>99</td>
</tr>
<tr>
<td>OUR PROPOSE</td>
<td>99.2</td>
<td>99.4</td>
<td>99.7</td>
</tr>
<tr>
<td>LPC WORDS</td>
<td>87.94</td>
<td>94.47</td>
<td>94.47</td>
</tr>
<tr>
<td>CLPC WORDS</td>
<td>90.45</td>
<td>97.99</td>
<td>98.99</td>
</tr>
<tr>
<td>MFCC WORDS</td>
<td>97.99</td>
<td>97.49</td>
<td>99.5</td>
</tr>
<tr>
<td>OUR PROPOSE</td>
<td>99.2</td>
<td>99.4</td>
<td>99.5</td>
</tr>
</tbody>
</table>

Table 1. Comparison between LPCs, CLPCs, MFCCs and our propose

<table>
<thead>
<tr>
<th>STATE NUMBER used in HMM/coefficients</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC SENTENCE</td>
<td>79</td>
<td>90.5</td>
<td>91.5</td>
</tr>
<tr>
<td>CLPC SENTENCE</td>
<td>93</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>MFCC SENTENCE</td>
<td>99</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>OUR PROPOSE</td>
<td>99.2</td>
<td>99.4</td>
<td>99.7</td>
</tr>
<tr>
<td>LPC WORDS</td>
<td>79.4</td>
<td>90.95</td>
<td>91.96</td>
</tr>
<tr>
<td>CLPC WORDS</td>
<td>93.47</td>
<td>98.5</td>
<td>99.5</td>
</tr>
<tr>
<td>MFCC WORDS</td>
<td>99.5</td>
<td>99.5</td>
<td>99.5</td>
</tr>
<tr>
<td>OUR PROPOSE</td>
<td>99.2</td>
<td>99.4</td>
<td>99.6</td>
</tr>
</tbody>
</table>

Table 3. Comparison between LPCs, CLPCs, MFCCs and our propose using DELTA, ACCELERATION, DELTA, AND THIRD DIFFERENTIAL coefficients.

4. Conclusions

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. For a long time a great diversity of models that describe the functionality of the ear have been proposed and they are implicit about this phenomenon. On the other hand, implanting cochlear is a new technology that joints with the models described, both using the place theory proposed by Von Bekesy. This proposal with the results obtained has been integrated into the ASRs task satisfactorily to reach a performance of 99.8%. This demonstrates the cochlea functionality for extracting information from the speech signal. Although the experiments were based over a digit task recognition task, the principal difference is the conformation of the filter bank, then the procedure is similar to obtain MFCC but a new distribution of the filter bank is considered which is based in cochlear operation. It can be probed with another database as TIMIT to try to testing the robustness of the results. At respect I consider that same good results can be obtained. Also, other cochlear models can be used but an equation that can relate frequency with distance need to be obtained, actually it is not exists.

5. Acknowledgements

I want to acknowledge The National Polytechnic Institute (IPN), also Center for Computing Research and especially SIP project number 20130985 for their support.
References
[42] http://htk.eng.cam.ac.uk/