

FRONT-END DESIGN BY USING AUDITORY MODELING IN SPEECH RECOGNITION

Jilei Tian, Kari Laurila, Ramalingam Hariharan, Imre Kiss

Speech and Audio Systems Laboratory
Nokia Research Center
P.O.Box 100, 33721 Tampere, Finland

ABSTRACT

A speech recognizer can commonly be divided into two main parts, front-end and back-end. The purpose of the front-end speech processing is to transform the original speech signal into a more suitable representation for recognition purposes. It is well known that the standard Mel-Frequency Cepstral Coefficients (MFCC) based front-end is quite vulnerable in the presence of environmental noise. In this paper, our aim is to achieve improved noise robustness by incorporating some of the key functions of the peripheral auditory system. By including nonlinear frequency scaling, intensity compression (loudness), short-term adaptation and firing rate of auditory neurons, we have obtained promising experimental results. We show that the proposed auditory front-end outperforms the conventional MFCC front-end, for the same feature vector dimension, in terms of recognition accuracy for speaker-dependent isolated-word recognition task, under different background noise conditions.

1. INTRODUCTION

It is very well known that the human auditory system is the most excellent speech recognizer. If such a computer-based speech recognition system could be designed that sufficiently reflects the process of auditory system, the resulting representations should be superior to representations based on non-biological criteria commonly used in computer speech recognition algorithms. The potential advantages of using auditory modeling for speech recognition task depend on how accurate the models are in mimicking human auditory system. Building such accurate models rely on the amount of knowledge we have about the auditory system. This knowledge is acquired by combining data that has been collected using psychophysical, physiological and auditory phenomena. Due to the extensive studies of auditory system, we now know quite a lot about the kinds of transformations that take place, at least at the peripheral level, and it has become feasible to build computational models that take these auditory phenomena into account. Different types of such speech signal auditory representations may make it easier to identify those features of the signal that are relevant for the speech recognition. In addition to the commonly used MFCC (Mel Frequency Cepstral Coefficients) front-end [7], many researchers have proposed alternative auditory approaches.

The perceptual linear predictive (PLP) technique proposed in [3] uses three concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum which is more consistent with human hearing. Even though PLP is computationally efficient and yields a low-dimensional representation of speech, some functions of auditory system like short-term adaptation are not considered. A joint synchrony/mean-rate auditory speech processing scheme proposed in [9] provided promising results in a case study but the combination with a commonly used HMM-based classifier was not very successful [4]. Cohen proposed another scheme in [1] but did not apply the method within the HMM framework. Ghiza [2] developed an ensemble interval histogram (EIH) model. In comparison with the commonly used MFCC front-end on an isolated word database in adverse conditions [4], the reduction of error rate was rather small with a high computational load. In addition, the research of auditory modeling has been widely carried out for purposes other than speech recognition [5].

In short, many researchers have shown that auditory modeling approach can lead to enhanced representations of speech signals. In this paper, we combine some previously proposed auditory functions and apply them to the front-end in order to obtain improved noise robustness.

2. AUDITORY FRONT-END

Since the human auditory system is not thoroughly understood and it can not be accurately modeled yet, in this paper, we only take some critical auditory functions relevant to speech recognition into account to build the auditory front-end. The basic idea is to incorporate nonlinear frequency scaling, intensity compression (loudness), short-term adaptation and firing rate of auditory neurons into the model.

$$\begin{cases} n(k) = \frac{r + n(k-1)}{1 + g_s + g_d + c \cdot s(k)} \\ f(k) = (g_s + c \cdot s(k)) \cdot n(k) \end{cases} \quad (3)$$

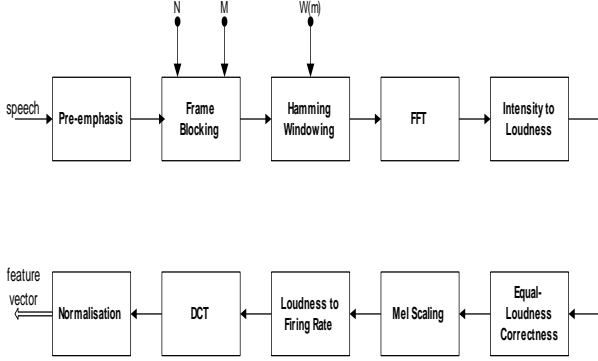


Figure 1. Block diagram of the auditory front-end.

A block diagram of the auditory front-end is given in Figure 1. The power spectrum of each frame is computed by applying FFT on windowed speech samples after pre-emphasis. Next we apply intensity to loudness conversion, also called as cubic root compression, i.e.: $\text{loudness} = (\text{intensity})^{1/3}$. This operation is an approximation to the power law of hearing and simulates the nonlinear relation between the intensity of sound and its

$$H(\omega) = 1.151 \cdot \sqrt{\frac{(\omega^2 + 144 \times 10^4) \omega^2}{(\omega^2 + 16 \times 10^4)(\omega^2 + 961 \times 10^4)}} \quad (1)$$

perceived loudness.

An approximation to the nonequal sensitivity of human hearing at different frequencies is given by equation (1). It simulates the sensitivity of hearing at about the 40 dB level [3].

A filter bank can be regarded as a crude model for the transduction of the basilar membrane in the human auditory system. A set of 24 bandpass filter banks, based on the mel scale is used to model the basilar membrane.

Next, the inner hair cell and attached auditory neuron are modeled as the transduction from loudness to firing rate. In the Schroeder-Hall model [8], “quanta” of an electrochemical agent are generated at a fixed average rate r . The probability of firing of an attached auditory neuron is directly proportional to the number of quanta currently existing and to the instantaneous input stimulus level $s(t)$ (the square root of loudness). The

$$\begin{cases} \frac{dn(t)}{dt} = r - (g_d + g_s + c \cdot s(t)) \cdot n(t), \\ f(t) = (g_s + c \cdot s(t)) \cdot n(t) \end{cases} \quad (2)$$

quanta are used up by producing spontaneous firings g_s and natural decay g_d without causing any firing. Thus equation (2) describes the number of quanta as a function of time and the instantaneous firing rate $f(t)$ of auditory neuron:

where $n(t)$ is the number of quanta at time instant t , r is the constant quanta generation rate, $s(t)$ is the square root of loudness of the input stimulus, c is the constant scale and $f(t)$ is

the instantaneous firing rate of auditory neuron attached to the inner hair cell. By transforming the above equation in discrete form, we have the following iterative form of the discrete nonlinear equation group:

SNR	clean	10 dB	0 dB	-10 dB
Aud FE	0.7123	0.6062	0.5016	0.3337
MFCC	1.2215	0.7445	0.4691	0.2380

By applying discrete cosine transform (DCT) on the firing rates from all the subchannels, we obtain 13 decorrelated features which form the feature vector for one frame.

3. EXPERIMENTS

The parameters (r , c , g_d , g_s) of the model (see equation (3)) are determined according to the relevant physiological data mentioned below [1][8]. First of all, firing rates in response to a tone burst can be simulated as the sum of two decaying exponentials. The time constant of fast adaptation is about 2 ms [11], which is too short to be significant in the frame-based features where frame shift is 10 ms. Another time constant is associated with the decreasing response to a stimulus which is a general characteristic of auditory neurons. It is around 30 ms (i.e.: 3 frames). When the stimulus is turned off, the firing rate recovers to the spontaneous rate with a time constant of 50 ms (i.e.: 5 frames).

3.1 Observations On TIMIT Database

The TIMIT database has been designed to provide speech data for the acquisition of acoustic-phonetic data knowledge and for the development and evaluation of automatic speech recognition systems including front-end. We have randomly picked an utterance *sal* (“she has your dark suit in greasy wash water all year”) in order to illustrate some simple comparisons between the MFCC and auditory front-ends.

The trajectories of the first components of the feature vectors generated by MFCC and auditory front-end are shown in Figure 2. It appears that the auditory front-end can capture dynamics better than the MFCC front-end. Specifically, the peaks at the transition portions of speech are emphasized and can be clearly observed. Though it looks like the auditory front-end might be capable of bringing some new information, it is not at all clear that it does so. In order to really know, we studied the separability between different phonemes at the feature level to base our conclusions on such more statistical measures. It might be that the peaks do not match well with the Gaussian density assumption and a reduced performance is achieved.

The separability of speech units in the feature space is a key indicator for evaluating the front-end. One of the J-measures defined in equation (4), was used for this purpose.

$$J = \frac{\text{tr}(\mathbf{B})}{\text{tr}(\mathbf{W})} \quad (4)$$

where matrix \mathbf{B} is the between-class covariance, or covariance of class means, and measures how close the speech classes are to each other. Matrix \mathbf{W} is the within-class covariance, or the average of the class covariances. We applied the J-measure as the phonetic separability indicator to the test set of TIMIT database containing 1680 sentences in both clean and noisy conditions. Table 1 gives the results for both MFCC and auditory front-end.

Table 1. Separability values (J-measure) of phonemes in the test set of TIMIT database for MFCC and auditory front-ends for clean and noisy speech.

Based on the separability values for phonemes, it can be seen that the auditory front-end can provide better discrimination ability in adverse conditions than MFCC front-end, but worse in cleaner environments.

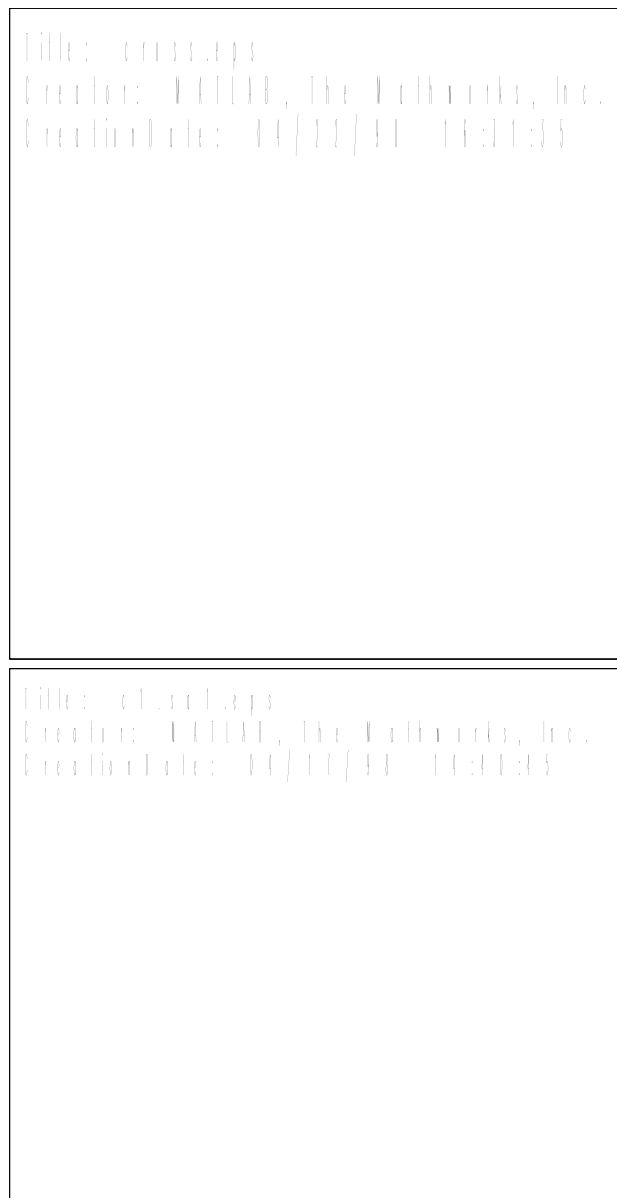


Figure 2. The feature trajectory of an utterance

While many current speech recognizers provide rather good recognition accuracy in noise-free conditions, their performance degrades rapidly when they are exposed to noisy environments.

In order to get noisy samples, noise from a Volkswagen car traveling at 115 km/h was recorded and further mixed with clean speech (again the utterance *saI* is used for illustration) to generate noisy speech for different signal-to-noise ratios (SNR).

Figure 3 compares the noise robustness of the MFCC and auditory front-ends. First, each feature was normalized by removing the mean and normalizing the variance to be one. With each SNR (10, 0 and -10 dB), the cross correlation was calculated between the clean and the corresponding noisy

speech to measure their similarity. Obviously, if the cross-correlation is low, the features are heavily corrupted by the noise and, if the cross-correlation is high, it means the features are noise robust.

It is clear that the features produced by auditory front-end (solid line) are more noise robust than the features produced by MFCC front-end (dashed line) in this case study. We can also see that the features c_1 and c_0 are less distorted among all features.

3.2 Isolated-Word Recognition Test

SNR	clean	5 dB	0 dB	-5 dB	-10 dB
Aud FE	99.09	97.12	93.64	84.55	58.98
MFCC	99.43	96.36	91.59	80.42	53.41

The final goal in any front-end development work is an improved speech recognition accuracy. Improvements in visual representation level or in a certain phonetic separability measure are practically worthless without noticeable difference in the back-end. We decided to test the auditory front-end in an isolated-word speaker-dependent recognition task. Reason for such a test decision was that we have a high performance name dialling engine which is very difficult to improve.

The test database contained 30 Finnish first names spoken by six male and two female speakers. The recordings were carried out in an office environment during three separate sessions (12 repetitions of each name overall).

Figure 3. The similarity of the auditory (bold) and MFCC-based (dashed line) feature vectors between clean and noisy speech for different SNRs.

SNR	clean	5 dB	0 dB	-5 dB	-10 dB
Aud FE	99.09	97.12	93.64	84.55	58.98
MFCC 26	99.73	97.58	94.58	86.55	63.11

Again, noise from a Volkswagen car traveling at 115 km/h was recorded and further mixed with clean speech to generate the noisy speech under certain signal-to-noise ratios (SNR).

Continuous Gaussian density left-to-right state duration constrained hidden Markov models (HMMs) with a global variance vector were estimated with a single training utterance [6]. Table 2 summarizes the results obtained with the auditory and MFCC front-ends. It should be mentioned here that the MFCC based front-end produced 13 cepstral coefficients including the energy value. It can be seen that the auditory front-end provides enhanced noise robustness, though with somewhat lower recognition accuracy in a clean environment.

Table 2. Recognition rates for MFCC and auditory front-ends, without normalization, for different noise conditions.

We have previously proposed so called feature vector normalization method to enhance noise robustness of MFCC features [10]. With this normalization, short-term means and

variances of each feature vector component are set to zero and one respectively, regardless of environment.

When this normalization is performed on the auditory features, we hope that the sharp peaks in the trajectory of each feature vector component are suppressed and that the features become more suitable to the HMM framework (that is, fit better to unimodal Gaussian densities). We also hope that the other advantages of the normalization method, mentioned above, are also present in the auditory modeling case, and not just in the MFCC case.

Table 3 summarizes the results obtained with the auditory and MFCC front-ends with the normalization block enabled (see Fig. 1). It can be seen that the auditory front-end still provides enhanced noise robustness. Unfortunately the clean environment remains problematic for the auditory method, having a poorer recognition accuracy than the MFCCs.

Table 3. Recognition rates for MFCC and auditory front-ends, with normalization, for different noise conditions

Time domain dynamics of speech can be incorporated into the MFCC by adding the so called delta coefficients that are normally calculated with linear regression to estimate instantaneous derivatives (delta) for cepstral coefficients. All calculated delta parameters are appended to the feature vector. We compared the recognition performance between auditory front-end and MFCC with delta information. In Table 4, it is shown that MFCC with dynamics performs slightly better than auditory front-end. However, it should be noted that the length of feature vector using MFCC front-end is double that of auditory front-end.

Table 4. Recognition rates for MFCC (with dynamic information) and auditory front-ends, with normalization, for different noise conditions.

4. DISCUSSION

The auditory front end proposed in this paper incorporates some relevant auditory functions related to the inner ear. It captures

SNR	clean	5 dB	0 dB	-5 dB	-10 dB
Aud FE	97.42	91.02	81.63	58.90	27.12
MFCC	98.94	86.25	68.90	37.99	13.49

dynamic features such as onsets and offsets, producing observable peaks at the transients. This negates the need for incorporating separate features representing dynamic information as in the case of MFCC front-end. In this paper, it is shown that the auditory front-end performs better than MFCC front-end in all conditions except in clean environment. In addition, we show that the normalization method proposed earlier for the MFCC front-end further improves the performance of the auditory front end also, especially in noisy environments. This is due to the fact that the normalization

procedure reduces the mismatch between training and testing environments. Also, applying the normalization to the auditory front-end can suppress somewhat the sharpness of the peaks and make the features more suitable to fit into the HMM scheme.

In order to make the auditory front-end presented in this paper useful in practice, more experiments and modification needs to be performed to make it work better even in clean conditions. Incorporation of other auditory phenomena should be considered to increase the performance. The complete auditory framework can be viewed as consisting of two parts:

1. the peripheral auditory system (PAS), consisting of the outer, middle and inner ear and
2. the central auditory system (CAS) which form the succeeding processing stages upto the brain.

Our current research has focused on modeling the first part of the auditory system. We will henceforth concentrate on trying to incorporate several phenomena associated with the central auditory system.

It is also important to ensure that the resulting auditory front end integrates well with the HMM framework. These preliminary results can be considered as a promising starting point for further research.

REFERENCES

- [1] Cohen J.R. "Application of an auditory model to speech recognition". *Journal of the Acoustical Society of America*, 85(6):2623-2629, 1989.
- [2] Ghitza O. "Auditory model and human performance in tasks related to speech coding and speech recognition". *IEEE Trans. Speech and Audio Processing*, 2(1):115-132, 1994.
- [3] Hermansky H. "Perceptual linear predictive (PLP) analysis of speech". *Journal of the Acoustical Society of America*, 87(4):1738-1752, 1990.
- [4] Jankowski C.R., Vo H.H., and Lippmann R.P. "A comparison of signal processing front ends for automatic word recognition". *IEEE Trans. Speech and Audio Processing*, 3(4):286-293, 1995.
- [5] Kates J.M. "A time-domain digital cochlear model". *IEEE Trans. Signal Processing*, 39(12):2573-2592, 1991.
- [6] Laurila K., "Noise robust speech recognition with state duration constraint". *Proceedings of International Conference of Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, pages 871-874.
- [7] Picone J.W. "Signal modeling techniques in speech recognition". *Proceedings of the IEEE*, 81(9):1215-1247, 1993.
- [8] Schroeder M.R., and Hall J.L. "A model for mechanical to neural transduction in the auditory receptor". *Journal of the Acoustical Society of America*, 55(5):1055-1060, 1974.
- [9] Seneff S. "A joint synchrony/mean-rate model of auditory speech processing". *Journal of Phonetics*, 16:55-76, 1988.
- [10] Viikki O., and Laurila K. "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization". *Proceeding of ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 1997, pages 107-110.
- [11] Yates G., and Robertson D. "Very rapid adaptation in auditory ganglion cells". *Proceeding of the Fifth International Symposium on Hearing*, Delft, The Netherlands, 1980, pages 200-205.