NOISE ROBUST TWO-STREAM AUDITORY FEATURE EXTRACTION METHOD FOR SPEECH RECOGNITION

Jilei Tian, Ramalingam Hariharan, Kari Laurila

Speech and Audio Systems Laboratory, Nokia Research Center

P.O. Box 100, 33721 Tampere, Finland

Email: {jilei.tian, ramalingam.hariharan, kari.laurila} research.nokia.com

ABSTRACT

Part of the problems in noise robust speech recognition can be attributed to poor acoustic modeling and use of inappropriate features. It is known that the human auditory system is superior to the best speech recognizer currently available. Hence, in this paper, we propose a new two-stream feature extractor that incorporates some of the key functions of the peripheral auditory subsystem. To enhance noise robustness, the input is divided into low-pass and high-pass channels to form so-called static and dynamic streams. These two streams are independently processed and recombined to produce a single stream, containing 13 feature vector components, with improved linguistic information. Speaker-dependent isolated-word recognition tests, using the proposed front-end, produced an average 39% and 17% error rate reductions, over all noisy environments, as compared to the standard Mel Frequency Cepstral Coefficient (MFCC) front-ends with 13 (statics only) and 26 (statics and deltas) feature vector components, respectively.

1. INTRODUCTION

It is widely acknowledged that the performance of the current state-of-the-art speech recognizers starts to drop drastically in noisy conditions. It is hence clear that new technological breakthroughs are required for a major performance improvement. In order to make significant improvements, we need to acquire more basic knowledge in the area of feature extraction. As we know, modern speech recognizers still perform much worse than humans both in clean and noisy environments [1]. Our work presented in this paper is based on the assumption that front-ends based on human auditory system should be superior to other feature extraction approaches. Modeling the complete human auditory system is, however, not possible since the system is only partially understood. Nevertheless, some parts of the system are known and can hence be utilized to improve the feature extraction unit.

The human auditory system has an amazing ability to separate and understand sounds. It is commonly believed that temporal information plays a key role in this ability, more important than the frame-based spectral representation that is traditionally utilized in ASR front-ends. The conventional methods for incorporating temporal information into speech features apply linear regression to a series of successive cepstral vectors to generate difference cepstra [2], namely the delta and delta-delta coefficients. Significant improvements in performance have been achieved by adding the difference cepstra. However, the dynamic information provided by the linear regression is rather limited and methods that provide enhanced temporal information have been proposed. The relative spectral (RASTA) technique proposed in [3] to enhance the temporal features was shown to increase the recognition performance with convolutional channel noise. RASTA did not, however, give significant improvement in our speaker-dependent isolated-word recognition tests in noisy car environments. We recently proposed a new auditory front-end based on short-term adaptation [4]. In this paper, we propose further improvements to our earlier model that utilizes the temporal information more effectively.

2. AUDITORY FEATURE EXTRACTION

Several feature extraction techniques inspired by the human auditory system have been proposed over the past few years. However, the human auditory system is not thoroughly understood and hence can not be accurately modeled. In this paper, we therefore take only the critical auditory functions relevant to speech recognition task 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.

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

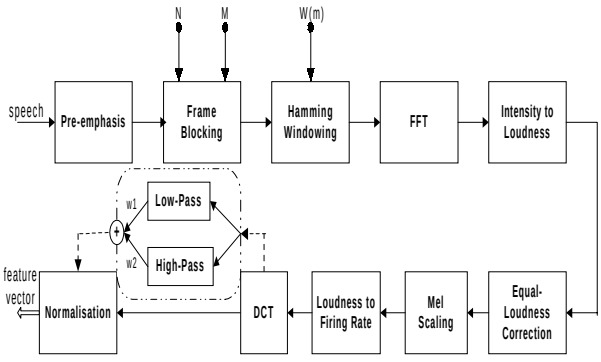

Fig. 1: Block diagram of the auditory front-end.

An approximation to the unequal sensitivity of human hearing at different frequencies is given by equation 1. This equal-loudness correction simulates the sensitivity of hearing at about the 40 dB level [3].

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

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

Next, the inner hair cell and attached auditory neuron is modeled as the transduction from loudness to firing rate. In the Schroeder-Hall model [5], "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 quanta are used up by producing spontaneous firings g_s and natural decay g_d without causing any firing. The number of quanta as a function of time and the instantaneous firing rate *f(t)* of auditory neuron are described by the equation:

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

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 a constant scaling factor and *f(t)* is the instantaneous firing rate of auditory neuron attached to the inner hair cell. Transforming the above equation into discrete form, we have the following iterative form of the discrete nonlinear equation group:

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

By applying discrete cosine transform (DCT) on the firing rates from all the sub-channels, we obtain 13 de-correlated features to form the feature vector.

A mismatch between training and testing environments can produce a severe degradation in the recognition performance. To reduce this mismatch, the normalization scheme proposed in [6] is carried out on the feature vector domain. With the normalization, short-term means and variances of each feature vector component are set to zero and one, respectively, regardless of environments.

2.1 Two-Stream Approach

The spectral trajectories of the feature vector components were studied in order to enhance the noise robustness of the previously proposed auditory front-end. Fig. 2 shows the ratio of the averaged spectra between the feature vector component trajectories of clean speech and car noise. A test set containing 110 sentences, spoken by seven male and four female speakers, was chosen from the TIMIT database. The ratio was computed by averaging across all the 13 feature vector components over all the utterances. We can observe that there is high local SNR in low frequency band, and the local SNR decreases as the frequency increases. Furthermore, it has been shown that the frequency content beyond a certain frequency value of feature vector component trajectory of the speech contains a significant amount of estimation error [7]. The front-end should be more noise robust if we can utilize the information according to the local SNR. The overall SNR could be increased by weighting the lower band more than the higher one. This approach also reduces the sharp peaks at the transitions produced by the short-term adaptation, resulting in parameter statistics that fit better into our HMM framework.

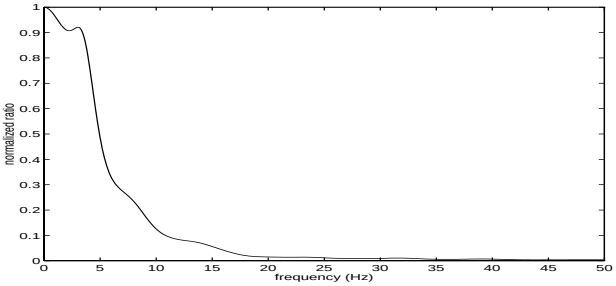

Fig. 2: Ratio of the averaged spectra of the feature vector component trajectories between clean speech and car noise normalized to be one at zero frequency.

In order to realize the weighting, the original feature stream, obtained from DCT, is split into low and high frequency channels. These two channels are later recombined by proper weighting and subjected to normalization to form the final feature vector.

We assume that the transfer functions of low-pass $H_l(z)$ and high-pass $H_h(z)$ filters are complementary, i.e.

$$H_l(z) + H_h(z) = 1 \qquad (4)$$

Therefore the equivalent transfer function *H(z)* of the recombined stream is given by the following equation:

$$\begin{aligned} H(z) &= w_l \cdot H_l(z) + w_h \cdot H_h(z) \\ &= 2\delta \cdot H_l(z) + (1 - \delta) \end{aligned}, \qquad (5)$$

where δ ($-1 \le \delta \le 1$) is defined as weighting factor, and the weights $w_l = 1 + \delta$ and $w_h = 1 - \delta$. Obviously, *H(z)* is low-pass, all-pass and high-pass filter when δ is 1, 0 and -1, respectively. Fig. 3 shows the amplitude response of the low-pass, high-pass and combined filter with δ=0.4. Based on our experiments, the optimum cut-off frequency was found to be around 5 Hz. Fig. 3 also shows the amplitude response of the conventional linear regression filter used to generate the delta coefficients. It is clear from the figure that the new high-pass filter contains more dynamic information than the conventional filter.

3. EXPERIMENTS

We tested the two-stream auditory front-end in an isolated-word speaker-dependent recognition task. The test database contained 30 confusable 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).

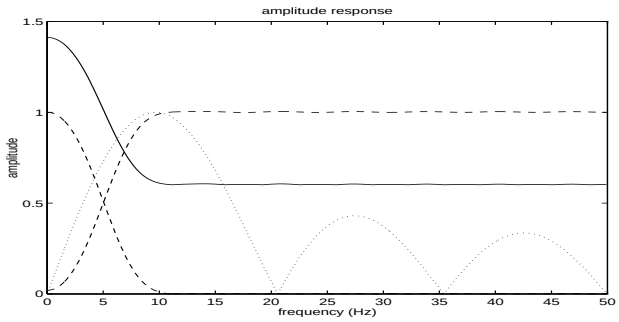

Fig. 3: Frequency response of the low-pass, high-pass (both seen as dashed line), combined filter (solid line, δ=0.4, cutoff frequency = 5 Hz) and the linear regression filter (dotted line, filter of length 7) .

Continuous Gaussian density left-to-right state duration constrained Hidden Markov Models (HMMs) [9] were estimated with a single training utterance, recorded in clean environment. Noise from a Volkswagen car traveling at 115 km/h was recorded and further mixed with clean speech to generate the noisy test utterances under certain signal-to-noise ratios (SNR).

The parameters (r, c, g_d ,g_s) of the auditory model (see equation 3) were determined according to the relevant physiological data mentioned in [5][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, which is too short to be significant in the frame-based features where frame shift is 10 ms. Hence we have not considered fast adaptation in our approach. Another time constant (around 30 ms) is associated with the decreasing response to a stimulus which is a general characteristic of auditory neurons**.** When the stimulus is turned off, the firing rate recovers to the spontaneous rate with a time constant of 50 ms. In order to optimize the weighting factor for the two-stream approach, recognition experiments were initially carried out at different SNRs. As shown in Fig. 4, the optimum weighting factor δ was found be around 0.4 ~ 0.6.

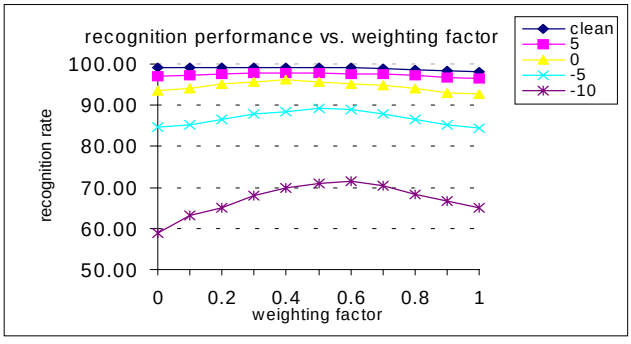

Fig. 4: Recognition rates at different SNRs for various values of the weighting factor δ.

3.1 Performance Evaluation

We randomly picked an utterance from TIMIT database *("she has your dark suit in greasy wash water all year")* in order to illustrate a simple comparison between the MFCC [2], previously proposed auditory front-end and the new two-stream auditory front-end. In order to verify the noise robustness of the different approaches, noisy speech was generated by adding car noise to clean speech at different SNRs. Each feature was normalized by removing the mean and normalizing the variance to be one. With each SNR, cross-correlation of the features was calculated between the clean and the corresponding noisy speech to measure their similarity. Fig. 5 compares the cross-correlation between the clean and noisy features for the three front-ends at different SNRs. Obviously, if the cross-correlation value is low, the feature is corrupted by the noise and if the cross-correlation value is high, it means that the feature is noise robust. It is clear that the features produced by two-stream auditory front-end are more noise robust than the features produced by the other front-ends in this case study, and the auditory front-end is more noise robust than the MFCC front-end.

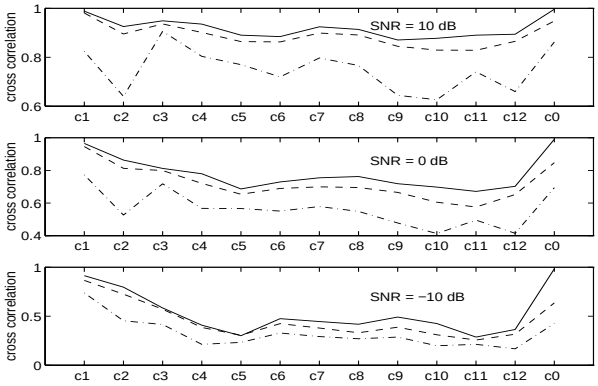

Fig. 5: Similarity measure of the features between clean and noisy speech at the different SNR conditions. The features were generated by the two-stream auditory (solid line), previously proposed auditory (dashed line) and MFCC (dash-dot line) front-ends.

Recognition tests were carried out between the previously proposed auditory front-end and the two-stream auditory front-end to compare their performance. Both approaches had a feature vector dimension of 13.

Fig. 6 shows the recognition results using the two front-ends. It is clearly seen that the two-stream auditory front-end outperforms the previously proposed auditory front-end. The average error rate reduction, over all noise conditions, was found to be around 27%.

We also compared the two-stream auditory front-end approach with standard MFCC front-end. Fig. 7 presents the results for the MFCC front-end with only static features (MFCC13) and also with both statics and deltas (MFCC26). Delta-delta coefficients were not used in these speaker dependent tests, as they produced worse results as compared to MFCC with statics and deltas. It can be clearly seen that the two-stream auditory approach outperforms the MFCC front-ends in all noisy

conditions. There is an average error rate reduction of 39% and 17% for the new approach over MFCC13 and MFCC26, respectively. However there seems to be a small decrease in the recognition performance in clean conditions

Finally, the superiority of the two-stream approach is demonstrated by comparing it to the previously proposed auditory front-end with delta features, computed using linear regression, thereby having a feature vector dimension of 26. It can be seen from Table 1 that the proposed two-stream auditory front-end, with a feature vector dimension of 13, produces a better recognition accuracy than the auditory front-end with delta features.

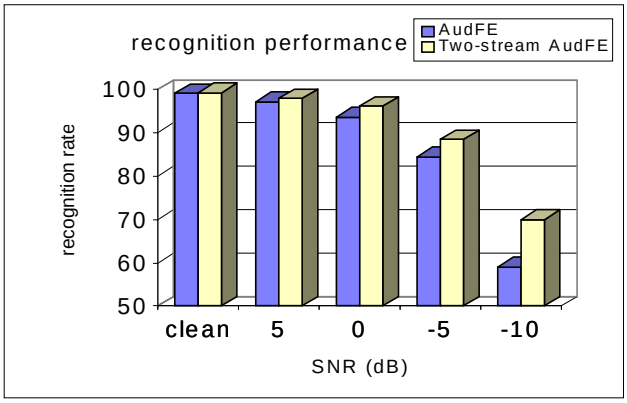

Fig. 6: Recognition rates at different SNRs using previously proposed auditory front-end (AudFE) and two-stream auditory front-end (two-stream AudFE).

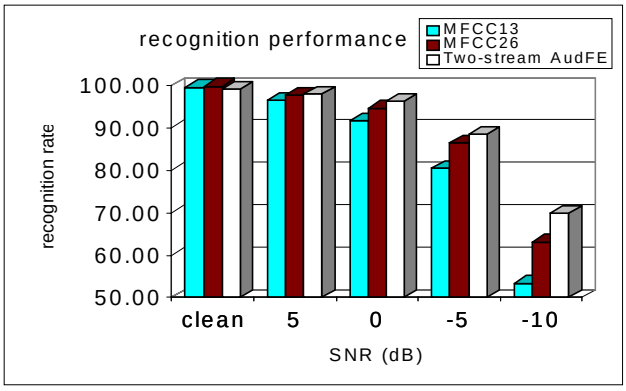

Fig. 7: Recognition rates at different SNRs using MFCC with only static features (MFCC13), MFCC with static and delta features (MFCC26) and the two-stream auditory front-end.

4. CONCLUSIONS

In this paper, we proposed a new noise robust two-stream auditory feature extraction method. Speaker-dependent isolated-word recognition tests performed using the new approach show that this front-end outperforms the conventional MFCC front-end in terms of recognition accuracy in all noisy environments. However there is a small decrease in the recognition accuracy in clean conditions, which requires further investigation. These results look promising enough to continue our study in this auditory domain to further improve our model and increase its performance.

Table 1: Recognition rates obtained with different front-ends at different SNRs.

SNR	MFCC13	MFCC26	AudFE	AudFE26	tsAudFE
clean	99.43	99.73	99.09	99.02	99.13
5	96.36	97.58	97.12	97.39	97.88
0	91.59	94.58	93.64	95.61	96.14
-5	80.42	86.55	84.55	88.60	88.45
-10	53.41	63.11	58.98	67.99	69.92
Ave.	84.24	88.31	86.67	89.72	90.30

REFERENCES

1. Jankowski, C.R., Vo, H.H., and Lippmann, R.P. "A comparison of signal processing front ends for automatic word recognition", *IEEE Trans. SAP*, 3(4):286-293, 1995.

2. Picone, J.W. "Signal modeling techniques in speech recognition", *Proc. of the IEEE*, 81(9):1215-1247, 1993.

3. Hermansky, H., et al., "RASTA-PLP speech analysis technique", *Proc. of ICASSP'92*, pp I-121-124, 1992.

4. Tian, J., Laurila, K., Hariharan, R., and Kiss, I. "Front-end design by using auditory modeling in speech recognition", *Proc. of NATO ASI on Computational Hearing*, pp 185-188, 1998.

5. Schroeder, M.R., and Hall, J.L. "A model for mechanical to neural transduction in the auditory receptor", *J. Acoustic. Soc. Amer.*, 55(5):1055-1060, 1974.

6. Viikki, O., and Laurila, K. "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization", *Proc. of ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp 107-110, 1997.

7. Nadeu, C., Paches-Leal, P., and Juang, B.H. "Filtering the time sequences of spectral parameters for speech recognition", *Speech Communication*, 22:315-332, 1997.

8. Cohen, J.R. "Application of an auditory model to speech recognition", *J. Acoustic. Soc. Amer.*, 85(6):2623-2629, 1989.

9. Laurila, K. "Noise robust speech recognition with state duration constraints", *Proc. of ICASSP'97*, pp II-871-874, 1997.