Blind Determination of the Signal to Noise Ratio of Speech Signals Based on Estimation Combination of Multiple Features

Russell Ondusko, Matthew Marbach, Andrew McClellan, Ravi P. Ramachandran, Linda M. Head
Rowan University
Correspondence: ravi@rowan.edu

Mark C. Huggins
Lockheed-Martin
Mark.Huggins@rl.af.mil

Brett Y. Smolenski
Research Associates for Defense Conversion
Brett.Smolenski@rl.af.mil

***Abstract*— A blind approach for estimating the signal to noise ratio (SNR) of a speech signal corrupted by additive noise is proposed. The method is based on a pattern recognition paradigm using various linear predictive based features, a vector quantizer classifier and estimation combination. Blind SNR estimation is very useful in speaker identification systems in which a confidence metric is determined along with the speaker identity. The confidence metric is partially based on the mismatch between the training and testing conditions of the speaker identification system and SNR estimation is very important in evaluating the degree of this mismatch. The aim is to correctly estimate SNR values from 0 to 30 dB, a range that is both practical and crucial for speaker identification systems. Additive white Gaussian noise and pink noise are investigated. The best feature for both white and pink noise is the vector of reflection coefficients which achieves an average SNR estimation error of 1.6 dB and 1.85 dB for white and pink noise respectively. Combining the estimates of 4 features lowers the error for white noise to 1.46 dB and for pink noise to 1.69 dB.**

I. Introduction

Consider a speech signal corrupted by additive noise that is statistically independent of the signal. This noisy signal is characterized by a signal to noise ratio (SNR) calculated over the entire duration of the signal. In this paper, a pattern recognition approach using various linear predictive (LP) [1] derived features is used to blindly estimate the SNR of the noisy speech signal. Blind estimation of the SNR is very useful in closed set speaker identification systems. The training of a speaker identification system involves the configuration of M models each representing a different speaker. During closed set testing, the features of an utterance are compared to the M models to render a decision of the speaker identity as being one of the M speakers [2][3]. Recent research has been done to develop techniques to calculate a confidence metric to accompany the decision of the speaker identity [4][5]. The confidence metric is calculated based on the mismatch between training and testing conditions, amount of training and testing data, and number of speakers (value of M). As M increases, there is usually more model overlap. The more the difference between the SNR of the training and testing speech, the more the mismatch between the two and the lower the confidence metric. An automatic and blind method of SNR estimation of the training and testing speech is an integral part of the technique of finding the confidence metric of a speaker identification system.

The method proposed for blind SNR estimation is based on a pattern recognition paradigm just like what is used for speaker identification. Features based on LP analysis that would not be robust to noise are highly useful candidates for SNR estimation as they show differences for varying noise levels. The overall system consists of four components, namely, (1) Linear predictive (LP) analysis, (2) Feature extraction for ensuring SNR discrimination, (3) Vector quantizer (VQ) classifier and decision logic for computing the SNR estimate and (4) Combination of the SNR estimates of the different features to get a final estimate. During training, a VQ codebook is trained for each distinct SNR value using feature vectors obtained from noisy speech corresponding to that particular SNR. During testing, the input to the system will be a noisy speech signal with an unknown SNR. After LP analysis and feature extraction, the set of feature vectors will be passed through each VQ codebook to get an overall distance for each codebook. Based on these distances, the output will be an estimated SNR value. A VQ classifier is trained separately for each feature and leads to an SNR estimate for each feature. A comparison of different LP based features is done with respect to the average absolute error between the actual and estimated SNR. The features considered [1][6][7] include the line spectral frequencies (LSFs), reflection coefficients (REFL), log area ratios (LAR), linear predictive cepstrum (CEP), adaptive component weighted cepstrum (ACW) and the postfilter cepstrum (PFL). The SNR estimates of the individual features are combined to get an even better estimate in that the average absolute error is further reduced.

II. Feature Extraction

Linear predictive analysis results in a stable all-pole model $1/A(z)$ of order p where

$$A(z) = 1 - \sum_{n=1}^{p} a(n) z^{-n} \tag{1}$$

The autocorrelation method of LP analysis gives rise to the predictor coefficients $a(n)$ and the REFL feature $refl(n)$ for $n = 1$ to p. The LAR feature is found as

$$lar(n) = \log\left[\frac{1 - refl(n)}{1 + refl(n)}\right] \tag{2}$$

for $n = 1$ to p. The LSF feature $lsf(n)$ are the angles (between 0 and π) of the alternating unit circle roots of $F(z)$ and $G(z)$ [1] where

$$\begin{aligned} F(z) &= A(z) + z^{-(p+1)} A(z^{-1}) \\ G(z) &= A(z) - z^{-(p+1)} A(z^{-1}) \end{aligned} \tag{3}$$

The predictor coefficients $a(n)$ are converted to the LP cepstrum $clp(n)$ $(n \geq 1)$ by an efficient recursive relation [1]

$$clp(n) = a(n) + \sum_{i=1}^{n-1} (\frac{i}{n}) clp(i) a(n-i) \tag{4}$$

Since $clp(n)$ is of infinite duration, the CEP feature vector of dimension p consists of the components $clp(1)$ to $clp(p)$ which

are the most significant due to the decay of the sequence with increasing n.

The first step in developing the ACW cepstrum [6] is to perform a partial fraction expansion of the LP function $1/A(z)$ to get

$$\frac{1}{A(z)} = \sum_{n=1}^{p} \frac{r_n}{1 - p_n z^{-1}} \tag{5}$$

where p_n are the poles of $A(z)$ and r_n are the corresponding residues. The variations in r_n were removed by forcing $r_n = 1$ for every n. Hence, the resulting transfer function is a pole-zero type of the form

$$\begin{aligned} \frac{N(z)}{A(z)} &= \sum_{n=1}^{p} \frac{1}{1 - p_n z^{-1}} \\ &= \frac{1}{A(z)} \sum_{n=1}^{p} \prod_{i=1 \neq n}^{p} (1 - p_n z^{-1}) \\ &= p\left[\frac{1 - \sum_{n=1}^{p-1} b(n) z^{-n}}{1 - \sum_{n=1}^{p} a(n) z^{-n}}\right] \end{aligned} \tag{6}$$

Applying the recursion in Eq. (4) to $b(n)$ and $a(n)$ results in two cepstrum sequences $cb(n)$ and $clp(n)$ respectively. The ACW cepstrum is $cacw(n) = clp(n) - cb(n)$ [6].

The postfilter is obtained from $A(z)$ and its transfer function is given by

$$H_{pfl}(z) = \frac{A(z/\beta)}{A(z/\alpha)} \tag{7}$$

where $0 < \beta < \alpha \leq 1$. The cepstrum of $H_{pfl}(z)$ is the PFL cepstrum which is equivalent to weighting the LP cepstrum as $cpfl(n) = clp(n)[\alpha^n - \beta^n]$ [7]. The ACW feature $cacw(n)$ and PFL feature $cpfl(n)$ are taken from $n = 1$ to p.

III. VQ Classifier and Decision Logic

A vector quantizer (VQ) classifier is used to generate a score for each candidate SNR value. During training, speech is corrupted by additive noise with a particular SNR and a corresponding set of feature vectors are computed. The feature vectors are used to design a VQ codebook for the particular SNR based on the Linde-Buzo-Gray algorithm [8]. The squared Euclidean distance is the distance measure. There will be N codebooks, one pertaining to each candidate SNR value.

During testing or score determination, a test noisy speech utterance of a particular SNR is converted to a set of test feature vectors. Consider a particular test feature vector. This is quantized by each of the N codebooks. The quantized vector is that which is closest with respect to the squared Euclidean distance measure to the test feature vector. Hence, N different distances are recorded, one for each codebook. This process is repeated for every test feature vector. The distances are accumulated over the entire set of feature vectors. This accumulated distance is the score for each codebook.

Two methods of implementing the decision logic are investigated. A hard decision approach estimates the SNR to correspond to the codebook which renders the smallest accumulated distance. This smallest distance is the best score. In the soft decision approach, the scores from a subset of the N codebooks are used to estimate the SNR. Consider the ith codebook trained for the value SNR(i) and rendering a score (accumulated distance) Score(i). Let Ind(i) denote the indicator function which equals 1 if codebook i is used for SNR computation. Otherwise, Ind(i) equals 0. The number of codebooks used which is also the number of times that Ind(i) equals 1 is denoted by C. A probability Prob(i) is derived from Score(i) by the equations

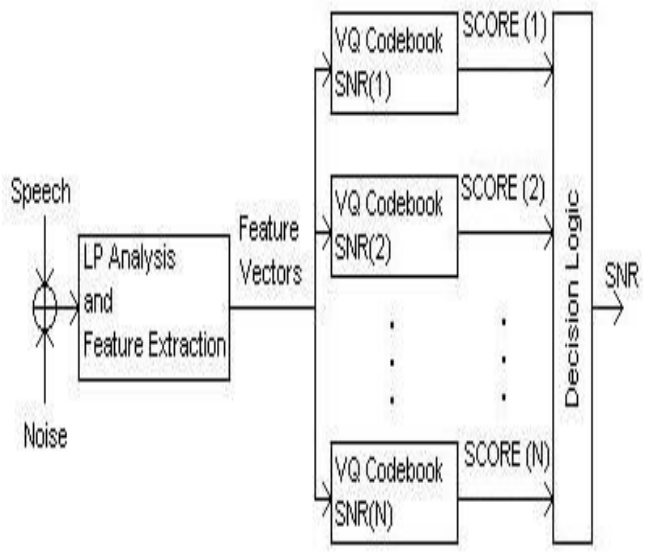

Fig. 1. Block diagram for SNR estimation using a single feature

$$\begin{aligned} \text{Total} &= \sum_{j=1}^{N} \text{Ind}(j)\text{Score}(j) \\ \text{Prob}(i) &= \text{Ind}(i)\left[\frac{\text{Total} - \text{Score}(i)}{(C-1)\text{Total}}\right] \end{aligned} \tag{8}$$

For the considered codebooks, smaller distances are converted to higher probabilities. If a codebook is not used, the probability assumes a value of 0. The probabilities add up to 1. The experiments revealed that using the three codebooks $(C = 3)$ with the smallest accumulated distances (best scores) led to good results. From the probabilities, the SNR is estimated as

$$\text{SNR} = \sum_{j=1}^{N} \text{Prob}(j)\text{SNR}(j) \tag{9}$$

For each test utterance, an absolute error between the true SNR and the estimated SNR is found. The performance measure is a mean value of this absolute error taken over the total number of test speech utterances. Figure 1 shows the block diagram for score and SNR determination.

IV. Estimation Combination

Using either hard or soft decision, 6 SNR estimates are found for each test speech utterance, one for each feature. A combination estimate is obtained by taking the mean, median and trimmed mean of all six or any subset of the individual feature SNR estimates. The aim is to see if all or a subset of the features contribute to a better final estimate. The trimmed mean is the mean of the estimates with the highest and lowest estimates not counted. It is only valid when three or more features are considered.

V. Experimental Protocol

Ten sentences from each of the 38 speakers from the New England dialect of the TIMIT database are used for the experiments. The speech in this database is clean and first downsampled from 16 kHz to 8 kHz. There are two training/testing scenarios, one in which white Gaussian noise is added and one in which pink noise is added. The noisy speech is preemphasized by using a nonrecursive filter $1-0.95z^{-1}$. For the LP analysis, the autocorrelation method [1] is used to get a 12th order LP polynomial $A(z)$. The LP analysis is done over frames of 30 ms duration. The overlap between frames is 20 ms. The LP coefficients are converted into 12 dimensional LSF, REFL, LAR, CEP, ACW and PFL feature vectors. For the PFL feature, $\alpha = 1$ and $\beta = 0.9$ (see Eq. (7)). The feature vectors are computed only in voiced frames that are selected based on energy thresholding. The VQ classifier (as described earlier) is trained using the 12 dimensional feature vectors. A separate classifier is used for each feature. A codebook of size 256 for each SNR value is designed using the Linde-Buzo-Gray algorithm [8]. The codebooks are designed for SNR values from 0 to 30 dB (inclusive) in 1 dB increments.

For each speaker in the database, there are 10 sentences. The first five are used for training the VQ classifier. The remaining five sentences are individually used for testing thereby giving 190 test cases. The roles of the training and testing speech are then reversed to get an additional 190 test cases bringing the total to 380. The goal is to correctly estimate SNR values between 0 and 30 dB (inclusive). This is a significant range for practical speaker identification systems. For each utterance, the absolute error is the absolute difference between the true SNR and the estimated SNR. For each SNR value tested, there are 380 utterances over which an average absolute error (AAE) is obtained. The AAE is found for each individual feature and for various combination estimates for both the hard and soft decision approaches.

VI. Results

An average absolute error (AAE) is computed for test speech having SNR values between 0 and 30 dB in 1 dB increments. There are a total of 31 AAE values and an average of these values result in an overall average absolute error (OAAE). Table I depicts these OAAE values for each of the six features when white Gaussian noise and pink noise is added. For the hard decision, there can be a zero error for a particular test speech utterance. The soft decision approach never gives a zero error but diminishes the OAAE when compared to the hard decision method. For white Gaussian noise, the three best features are the LSF, REFL and LAR features. The CEP feature shows only a slightly higher OAAE. For pink noise, the best features are the REFL and LAR features with the LSF feature showing a slightly higher OAAE.

Combination estimates using all possible subsets of the six features were attempted. Table II and Table III depict the OAAE values for the best combination estimates. The best subsets involving 2,3,4 and 5 features along with the fusion of all 6 festures are presented. When using a subset of 2 features, the trimmed mean has no meaning. When using a subset of 3 or 4 features, the trimmed mean is the same as the median.

Generally, the same feature combinations do well for both white and pink noise. Going from using just one feature to a combination of two features gives the maximum improvement in the OAAE. Then, increasing the number of features used to three and four gives a slight improvement. The OAAE does not decrease further after four features are used. For both white and pink noise, it is best to use a soft decision approach and a four feature combination estimate, namely, LSF/REFL/LAR/ACW or LSF/REFL/LAR/PFL. Figures 2 and 3 show the average absolute error (AAE) as a function of the SNR of the test speech for a single feature and the LSF/REFL/LAR/PFL feature combination. When four features are used, the AAE improves over that of a single feature for a wide range of SNRs and is about the same for very low and very high SNRs.

	White Noise		Pink Noise	
Feature	Hard Decision	Soft Decision	Hard Decision	Soft Decision
LSF	1.76	1.61	2.18	1.94
REFL	1.84	1.62	2.17	1.85
LAR	1.83	1.59	2.22	1.89
CEP	1.85	1.68	2.24	2.01
ACW	2.09	1.85	2.49	2.14
PFL	2.01	1.78	2.44	2.09

TABLE I
HARD AND SOFT DECISION OAAE VALUES (IN DB) FOR WHITE GAUSSIAN AND PINK NOISE

Features	Hard Decision	Soft Decision
LSF REFL	1.60 1.60 -	1.51 1.51 -
LSF LAR	1.60 1.60 -	1.50 1.50 -
LSF REFL LAR	1.55 1.59 1.59	1.49 1.49 1.49
LSF LAR ACW	1.55 1.59 1.59	1.49 1.49 1.49
LSF LAR PFL	1.55 1.60 1.60	1.48 1.50 1.50
LSF REFL LAR ACW	1.51 1.54 1.54	1.47 1.47 1.47
LSF REFL LAR PFL	1.52 1.55 1.55	1.46 1.47 1.47
LSF REFL LAR CEP ACW	1.50 1.54 1.53	1.47 1.48 1.47
LSF REFL LAR CEP PFL	1.50 1.55 1.54	1.47 1.48 1.47
LSF REFL LAR ACW PFL	1.51 1.54 1.53	1.48 1.48 1.47
All Six	1.50 1.52 1.51	1.47 1.47 1.47

TABLE II
HARD AND SOFT DECISION OAAE VALUES (IN DB) FOR THE BEST COMBINATION ESTIMATES USING 2,3,4,5 AND ALL 6 FEATURES FOR THE CASE OF WHITE GAUSSIAN NOISE. THE THREE OAAE VALUES REFER TO THE MEAN, MEDIAN AND TRIMMED MEAN COMBINATION ESTIMATES. WHEN TWO FEATURES ARE USED, THE TRIMMED MEAN IS NOT CALCULATED.

VII. Summary and Conclusions

The VQ based pattern recognition approach to blind SNR estimation has given very good results. It is important to combine the soft decision estimates of 4 features to get the lowest possible OAAE of 1.46 dB for white noise and 1.69 dB for pink noise.

VIII. Acknowledgement

This work was supported by the U.S. Air Force Research Laboratory, Rome NY, under contracts FA8750-05-C-0029 and F30602-03-C-0067.

IX. References

1) T. F. Quatieri, *Discrete Time Speech Signal Processing Principles and Practice* Prentice Hall PTR, 2002.
2) J. P. Campbell, "Speaker recognition: A tutorial", *Proc. IEEE*, vol. 85, pp. 1437–1462, September 1997.
3) H. A. Murthy, F. Beaufays, L. P. Heck and M. Weintraub, "Robust text-independent speaker identification over telephone channels", *IEEE Trans. on Speech and Audio Proc.*, vol. 7, pp. 554–568, Sept. 1999.

Features	Hard Decision	Soft Decision
LSF REFL	1.89 1.89 -	1.74 1.74 -
LSF LAR	1.90 1.90 -	1.77 1.77 -
LSF REFL LAR	1.81 1.88 1.88	1.71 1.74 1.74
LSF REFL ACW	1.82 1.88 1.88	1.72 1.75 1.75
LSF LAR ACW	1.82 1.90 1.90	1.72 1.76 1.76
LSF LAR PFL	1.83 1.91 1.91	1.73 1.76 1.76
LSF REFL LAR ACW	1.76 1.80 1.80	1.69 1.70 1.70
LSF REFL LAR PFL	1.76 1.80 1.80	1.69 1.70 1.70
LSF REFL LAR CEP ACW	1.74 1.78 1.77	1.69 1.70 1.69
LSF REFL LAR CEP PFL	1.75 1.80 1.77	1.69 1.71 1.69
LSF REFL LAR ACW PFL	1.75 1.80 1.77	1.70 1.71 1.70
All Six	1.73 1.76 1.75	1.69 1.70 1.69

TABLE III

HARD AND SOFT DECISION OAAE VALUES (IN DB) FOR THE BEST COMBINATION ESTIMATES USING 2,3,4,5 AND ALL 6 FEATURES FOR THE CASE OF PINK NOISE. THE THREE OAAE VALUES REFER TO THE MEAN, MEDIAN AND TRIMMED MEAN COMBINATION ESTIMATES. WHEN TWO FEATURES ARE USED, THE TRIMMED MEAN IS NOT CALCULATED.

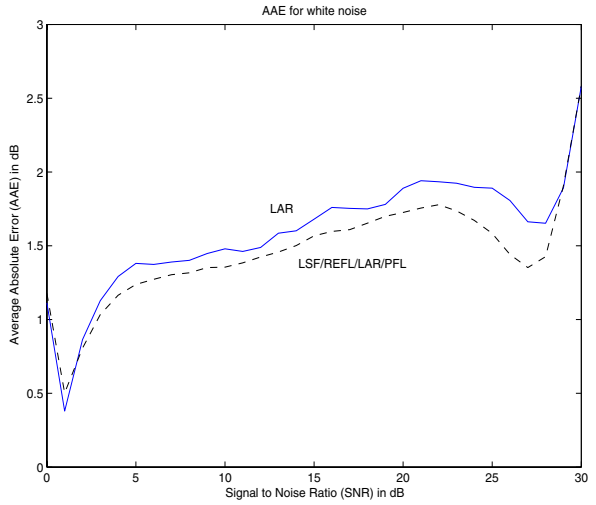

Fig. 2. Average absolute error versus SNR of the test speech for the LAR feature and the LSF/REFL/LAR/PFL combination for white noise

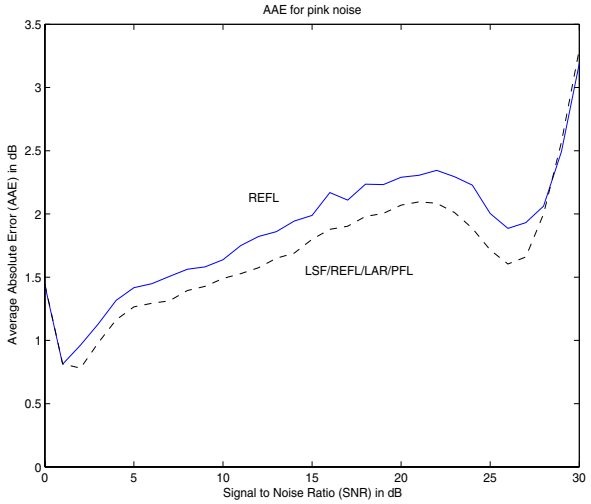

Fig. 3. Average absolute error versus SNR of the test speech for the REFL feature and the LSF/REFL/LAR/PFL combination for pink noise

4) M. C. Huggins and J. J. Grieco, "Confidence Metrics For Speaker Identification", *Int. Conf. on Spoken Language Processing*, Denver, Colorado, 2002.
5) M. C. Huggins and J. J. Grieco, "Speaker Identification Confidence Metrics For Heterogeneous Model Spaces", *Proc. of the 8th World Multiconference on Systemics, Cybernetics and Informatics*, Orlando, Florida, pp. 440–443, July 2004.
6) K. T. Assaleh and R. J. Mammone, "New LP-derived features for speaker identification", *IEEE Trans. on Speech and Audio Proc.*, vol. 2, pp. 630–638, Oct. 1994.
7) M. S. Zilovic, R. P. Ramachandran and R. J. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions", *IEEE Trans. on Speech and Audio Proc.*. vol. 6, pp. 260–267, May 1998.
8) Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. on Comm.*, vol. COM-28, pp. 84–95, Jan. 1980.