Optimization of Bayesian Adaptation in Continuous Robust Speech Recognition

Tatiane Melo Vital
Instituto Nacional de Telecomunicações - Inatel
P.O. Box 05 - 37540-000
Santa Rita do Sapucaí - MG - Brazil
tati.mv7@gmail.com

Carlos Alberto Ynoguti
Instituto Nacional de Telecomunicações - Inatel
P.O. Box 05 - 37540-000
Santa Rita do Sapucaí - MG - Brazil
ynoguti@inatel.br

***Abstract*— The main contribution of the present work is to provide an algorithm based on parametric adjustment (using a logistic curve) that returns good adaptation coefficients for Automatic Speech Recognition Systems which employs Maximum a Posteriori criteria and multi-style training.**

***Index Terms*— Automatic Speech Recognition, Maximum a Posteriori Adaptation, Multi-Style Training, Parametric Adjustment.**

I. Introduction

It is widely known that ASR (Automatic Speech Recognition) systems performance degrades when operating under noisy conditions [1], and one reason for this fact is the mismatch between the training and the testing acoustic conditions [2].

There are several approaches to minimize the effects of background disturbance even in unknown noisy conditions [3]. They can be divided in three classes:

- the first one is applied before acoustic modeling in front-end signal preprocessing. PLPs or MFCCs helps to minimize the effect of speaker variability. In front-end signal processing, noise suppression methods such as Subtraction Spectral (SS), Wiener filtering and Minimum Mean Square Error (MMSE) estimation are effective to reduce the intensity of noise in the speech;
- the second one take into account methods that acts in the modeling phase. In this case, clean speech is used on training to ensure the high quality of the final models of speech. Then, these models can be transformed according to noise present during recognition task. This category comprehends: techniques which combine background noise with speech, i.e multi-condition models, or with acoustic models such as parallel model combination (PMC);
- the last one approach includes methods which use noisy speech data to adapt acoustic models for a specific background condition by retraining the clean speech models or simply by some transformation as maximum likelihood linear regression (MLLR) or MAP adaptation.

To overcome this problem, in [4] an approach using multi-style condition training [5] followed by a Maximum a Posteriori (MAP) adaptation [6] was proposed. This method can be summarized as follows:

- in the first stage (multi-style training), a HMM is trained using utterances corrupted by several noise types available on AURORA database [7], at SNRs of 15 dB and 20 dB (this choice of SNRs is based on experimental results [4]). This stage provided a 6.89 % gain in WA (word accuracy) for noisy utterances recognition when compared with a system trained only with clean speech.
- in the second stage, a MAP adaptation was performed to fine tune the system for the actual noise type and SNR that is being experienced by the recognizer. An additional 1.74 % gain in WA was obtained, and thus, the overall gain with these two techniques is 8.63 % over the baseline system.

The purpose of this work is to allow the selection of good adaptation coefficient values without increasing the computational cost. The next sections are structured as follows: in Section II, the robustness techniques used are described. Section III presents the proposed method. In Section IV, the experimental setup used for recognition tests is demonstrated and Section V shows the test results. Finally, Section VI brings the final conclusion for present work.

II. Noise Robustness

ASR in noisy environments has been a challenging issue because the presence of noise decreases the accuracy of these systems. There are several approaches to reduce the influence of background noise on the performance of ASR systems. This work evaluates the combination of multi-style training and MAP estimation, techniques which are presented in the next subsections, to overcome effects caused by different noise types and levels.

To find the optimal value of adaptation coefficient used in the MAP approach is a computationally expensive procedure, since it involves a grid search. The main contribution of this work is to find a way to do this automatically, according noise type and level.

A. Multi-Style Training

Aiming to add robustness to Hidden Markov Models (HMM) to several environmental variabilities, channel distortion, reverberation among other unwanted effects, this technique relies on the availability of a collecting speech database in a real environment. However, the construction of a database that reflects all situations of day to day is infeasible and impractical given the great variability of environmental adverse conditions.

The multi-style or multi-condition training employs utterances artificially corrupted by different noise type and levels in the training stage in order to minimize the performance drop of ASR systems operating in noisy environments [8].

Different approaches can be used in this method: the system can be trained for a particular noise type and level according to environmental condition, or with different levels of a specific noise type, or even, with different noise types and levels. According to [4], this present work employs the last approach.

B. Maximum a Posteriori Adaptation

In the MAP approach, the models are adapted with estimated statistics (mixture weight, mean and variance) of the background noise. Generally, Bayesian adaptation returns good word accuracy because it provides the modeling of the uncertainty caused by noisy environmental statistics [9].

The main motivation to employ Bayesian adaptation of the canonical model generated from multi-style training is to achieve a superior performance by adapting the models for a specific noise type and level in the recognition step.

A canonical model is a HMM generated in the training phase using noisy or clean utterances of several speakers. Then noise statistics from environment are used to adapt these models. The hypothesized speech model is derived by adapting the parameters of canonical model and a form of Bayesian adaptation [6].

The adaptation equations for these parameters are described as follows. Given a noise sample and training vectors from the hypothesized speech, $X = x_1, x_2, ..., x_T$, the probabilistic alignment of the noise into the canonical model is given by:

$$\Pr(i|x_t) = \frac{\omega_i p_i(x_t)}{\sum_{j=1}^{M} \omega_j p_j(x_t)} \tag{1}$$

where M is the number of unimodal Gaussian densities, ω is the mixture weight and p is the probability density function.

Then, $Pr(i|x_t)$ and x_T are used to determinate noisy statistical parameters weight (n_i), mean ($E_i(x)$) and variance ($E_i(x^2)$), as described below:

$$n_i = \sum_{t=1}^{T} Pr(i|x_t) \tag{2}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t \tag{3}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t^2 \tag{4}$$

Finally, these estimated statistics of background noise are used to adapt the canonical models generating a new model. The adaptation equations for these parameters are:

$$\hat{\omega}_i = [\alpha_i^{\omega} n_i/T + (1 - \alpha_i^{\omega})\omega_i]\,\gamma \tag{5}$$

$$\hat{\mu}_i = \alpha_i^{m} E_i(x) + (1 - \alpha_i^{m})\mu_i \tag{6}$$

$$\hat{\sigma}_i^2 = \alpha_i^{\nu} E_i(x^2) + (1 - \alpha_i^{\nu})(\sigma_i^2 + \mu_i^2) - \mu_i^2 \tag{7}$$

where:

- ω_i, μ_i and σ_i^2 are the mixture weights, means and variances of the multi-style trained system;
- $\hat{\omega}_i$, $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are the mixture weights, means and variances after adaptation and
- n_i, $E_i(x)$ and $E_i(x^2)$ are the noise statistics.

The adaptation coefficients α_i^{ω}, α_i^{m} and α_i^{ν} can assume values in the $[0, 1]$ interval and control the balance between old and new estimates for the weights, means and variances, respectively.

It seems reasonable that a good choice of these parameters depends on the noise type and intensity, since higher values favor the noise estimates whereas lower values tend to preserve the original ones.

It is possible to employ different coefficient values to adapt weights, means and variances. However, this approach provides a small gain compared to use solely a single value (α_i^{ω}=α_i^{m}=α_i^{ν}) when adapting them. Due this, this work uses a single adjustment coefficient for all parameter as demonstrated in [6].

A scan by choosing the appropriate value of alpha is an arduous task due to the high computational cost. Section III describes a method for good choice for this parameter.

III. Modeling for Alpha Coefficient

The present work proposes an algorithm which provides an adaptation coefficient alpha value based on statistical characteristics of the background noise focusing the trade-off between system performance and maximum Word Accuracy. It suggests a new empirical approximation for modeling the relation between SNR and adaptation coefficient for each type of noise. The aim is to predict the α value for a wide range of noise levels according with various environmental conditions based on results from recognition tests for a limited range.

The aim of proposed method is to achieve a value for this parameter that provides a good Word Accuracy to a given level and type of each noise. Its the main effort is not to achieve the maximum system performance but to provide a value that returns a gain compared to the baseline.

The algorithm is outlied below:

- for each value of SNR, perform a grid search for the best values of alpha in the [0,1] interval and record the ones that led to performance improvement (in our experiments, we used 0 dB, 5 dB, 10 dB, 15 dB and 20 dB).
- for a given SNR, there are several values of α that lead to performance improvement, but we keep only one. After several tests, we decided to choose a weighted average, given by:

$$\alpha' = \frac{\sum_i WA(i) \times \alpha(i)}{\sum_i WA(i)} \quad (8)$$

where $WA(i)$ is the word accuracy obtained by using the value $\alpha(i)$, and α' is the chosen value for the α parameter. Therefore, after this step, there is a single value α' for each value of SNR.

- By looking at the curves generated by several tests, we observed that they resembled the format of a logistic curve. Thus, as a final step, a logistic curve (Equation 5) with three free parameters was adjusted to the experimental points [10].

$$f(x) = \frac{1}{1 + e^{b-ax}} - c \quad (9)$$

where x is the noise level and $f(x)$ is the adaptation coefficient. The configuration parameters a, b and c were obtained by an approximation using the test results. They can be interpreted as follows:

  - a parameter determines the slope of the logistic curve. As smaller its value, steeper is the curve;
  - b parameter controls the horizontal offset. If its value decreases, the curve is shifted to the right. Otherwise, it is shifted to left and
  - c parameter is the offset, allowing the vertical adjustment. If its value increases, the curve is moved down. Otherwise, it is shifted up.

  This is an important step because it allows the choice of α values for SNRs different than the ones used to generate the curve.

In the next section is presented the experimental setup used in the recognition tests.

IV. Experimental Setup

In this section, the database and speech recognition engine are described.

A. Database

Experiments were performed using a 40 speakers (20 male and 20 female) clean speech database [11]. Each speaker recorded 40 phonetically balanced utterances in Brazilian Portuguese which were drawn from [12]. The corpus has 1600 sentences comprising 694 different words and it was subdivided in two groups: training corpus (1200 utterances) and testing corpus (400 utterances).

The recordings were performed in low noise environment at 11,025 kHz sample rate and 16-bit coded. The sampling frequency was lowered to 8 kHz to make them compatible with the AURORA database.

The original database was artificially corrupted by the noises (airport, babble, car, exhibition, restaurant, street, subway and train) of the Aurora Project database [7]. For the training material, for each clean sentence, 2 new versions were created combining each noise type at levels 15 and 20 dB as proposed by [4]. For each clean utterance of testing corpus, 5 new versions were created adding each noise type at levels 0, 5, 10, 15 and 20 dB. Therefore, training and testing database have 19200 and 16000 corrupted utterances, respectively.

B. Speech recognition engine

To test our ideas, a continuous density HMM based speech recognition engine developed by [11] was used. It uses context independent phones as fundamental units where each of them is modeled as a 3 state Markov chain as shown in Figure 1 and the One Pass search algorithm [13]. A mixture of 10 multidimensional Gaussian distributions with diagonal covariance matrix was used in each state.

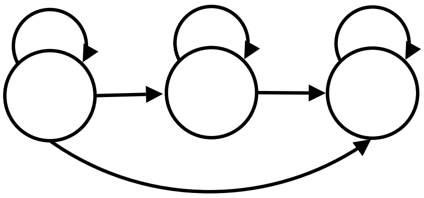

Fig. 1. Markov chain for each phone model

For acoustic parameters, it was used 12 mel-cepstral coefficients together with their first and second derivatives leading to feature vectors of dimension 36. Finally, to improve the system performance, a bigram language model was applied.

V. Experimental Results

This section describes tests and their results. It demonstrates by the experimental results that the proposed procedure presented in Section III led to a performance improvement, although not to the best possible one.

The baseline proposed is an ASR system trained using multi-style approach and uncleaned utterances in the recognition phase. For training, all noise types available on AURORA database at SNR level 15 and 20 db were used. Its performance was taken as reference WA value as can be seen on subsequent tables.

To evaluate the improvement of the recognition performance using multi-condition training and MAP adaptation combined together, the HMM obtained in the baseline system was adapted for each noise type, generating new 8 models. After that, recognition tests were performed for each one using corrupted utterances by the same noise type than adaptation stage.

Their results were computed and analyzed. The maximum WA, adaptation coefficient for maximum WA and weighted alpha for airport, babble, car, exhibition, restaurant, street, subway and train noise, are shown in Tables I, II, III, IV, V, VI, VII and VIII, respectively.

From Tables I, II, IV, VI, and VIII, it is possible to see that for some SNRs it doesn't exist a weighted α value because all tests returned a maximum WA that is smaller than reference value. For these cases, MAP adaptation introduced a little performance drop or did not provide gain comparing to the

TABLE I

AIRPORT TEST RESULTS.

SNR (dB)	Reference WA (%)	Maximum WA (%)	α for maximum WA	weighted α	WA for weighted α (%)
0	3.2	7.3	0.6500	0.3387	6.3
5	25.4	32.1	0.4500	0.2779	31.1
10	62.9	66.2	0.1500	0.1950	65.6
15	78.0	77.8	0.0100	–	–
20	78.0	77.8	0.0100	–	–

TABLE II

BABBLE TEST RESULTS.

SNR (dB)	Reference WA (%)	Maximum WA (%)	α for maximum WA	weighted α	WA for weighted α (%)
0	4.6	5.9	0.1500	0.2172	5.8
5	32.2	37.4	0.2000	0.2139	37.1
10	66.3	66.7	0.1500	0.1500	66.2
15	77.1	76.9	0.0100	–	–
20	77.1	76.9	0.0100	–	–

TABLE III

CAR TEST RESULTS.

SNR (dB)	Reference WA (%)	Maximum WA (%)	α for maximum WA	weighted α	WA for weighted α (%)
0	5.3	7.1	0.3000	0.2167	6.0
5	31.7	37.3	0.3500	0.2195	36.3
10	65.8	67.7	0.1500	0.1017	67.5
15	76.5	76.6	0.0400	0.0350	76.3
20	76.5	76.6	0.0400	0.0350	76.3

TABLE IV

EXHIBITION TEST RESULTS.

SNR (dB)	Reference WA (%)	Maximum WA (%)	α for maximum WA	weighted α	WA for weighted α (%)
0	0.2	1.8	0.4000	0.2310	1.0
5	16.4	23.5	0.0900	0.2283	23.1
10	58.2	58.7	0.1500	0.1267	58.0
15	75.5	75.5	0.0200	–	–
20	75.5	75.5	0.0200	–	–

TABLE V

RESTAURANT TEST RESULTS.

SNR (dB)	Reference WA (%)	Maximum WA (%)	α for maximum WA	weighted α	WA for weighted α (%)
0	13.8	15.4	0.2500	0.1751	14.8
5	4.4	9.1	0.3500	0.2371	8.5
10	35.5	41.6	0.3500	0.1976	41.1
15	69.2	69.7	0.0300	0.0325	69.7
20	69.2	69.7	0.0300	0.0325	69.7

TABLE VI

STREET TEST RESULTS.

SNR (dB)	Reference WA (%)	Maximum WA (%)	α for maximum WA	weighted α	WA for weighted α (%)
0	15.6	18.6	0.2000	0.1043	18.4
5	56.1	57.1	0.0700	0.0600	56.3
10	73.6	73.6	0.0200	–	–
15	75.8	77.1	0.0200	0.0250	76.8
20	75.8	77.1	0.0200	0.0250	76.8

TABLE VII

SUBWAY TEST RESULTS.

SNR (dB)	Reference WA (%)	Maximum WA (%)	α for maximum WA	weighted α	WA for weighted α (%)
0	-0.3	3.4	0.4000	0.3186	2.7
5	18.4	25.8	0.3000	0.2594	25.6
10	57.9	61.8	0.0500	0.1193	61.7
15	73.0	73.4	0.0800	0.0650	72.9
20	73.0	73.4	0.0800	0.0650	72.9

TABLE VIII

TRAIN TEST RESULTS.

SNR (dB)	Reference WA (%)	Maximum WA (%)	α for maximum WA	weighted α	WA for weighted α (%)
0	7.6	12.6	0.4500	0.3174	12.1
5	34.0	40.7	0.4500	0.2398	38.8
10	69.0	70.0	0.0400	0.0749	69.9
15	78.1	78.1	0.0100	–	–
20	78.1	78.1	0.0100	–	–

baseline. However, the results shows that this approach has a good cost benefit relationship.

Analyzing and computing these results, given the similarity between SNR x adaption coefficient and logistic curve, the parametric adjustment used to model this relation is based on logistic function. Thus, the logistic curve was traced for each noise type as shown in Figures 2, 3, 4, 5, 6, 7, 8 and 9. Table IX shows the configuration parameters: a, b and c that describes their behaviors.

TABLE IX

COEFFICIENTS FOR LOGISTIC CURVE.

Noise	a	b	c
airport	-0.104175	-0.810155	0.186922
babble	-0.319246	1.338508	-0.009345
car	-0.442008	1.445377	-0.025966
exhibition	-0.124371	1.049212	0.028005
restaurant	-0.148612	1.701653	-0.020750
street	-1.264809	2.320370	-0.014900
subway	-0.131225	1.000045	-0.049691
train	-0.299248	0.882365	0.000910

An important consequence is that these logistic curves may be used to predict a good adjustment coefficient value for a given SNR for each noise type. Moreover, the computational

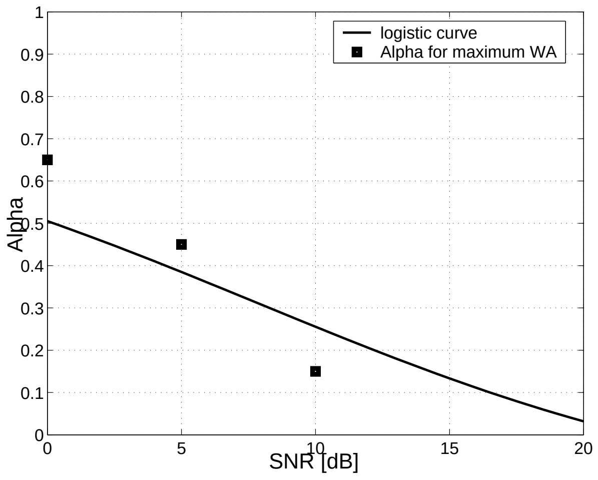

Fig. 2. Logistic curve for recognition using airport noise

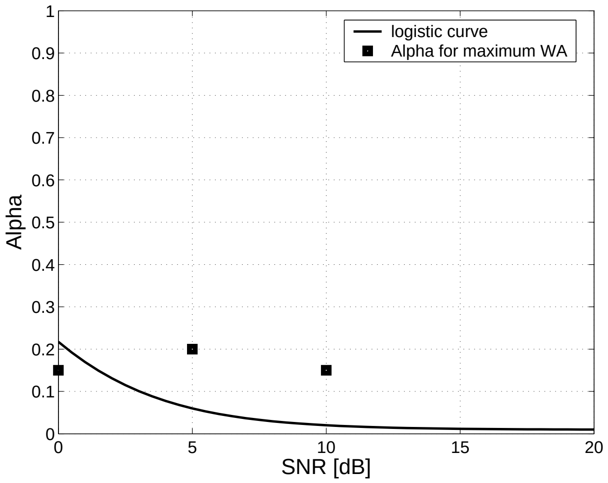

Fig. 3. Logistic curve for recognition using babble noise

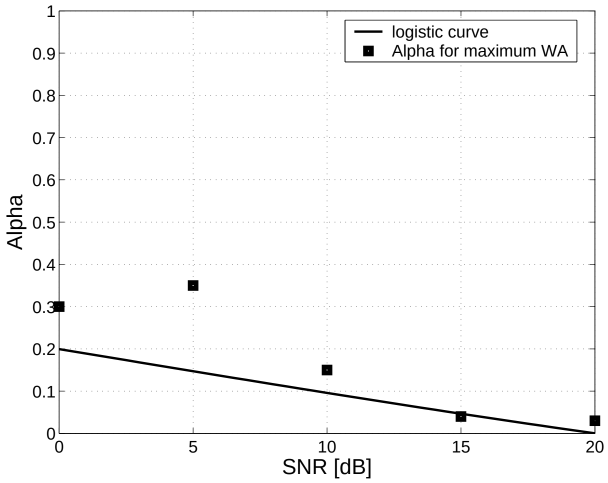

Fig. 4. Logistic curve for recognition using car noise

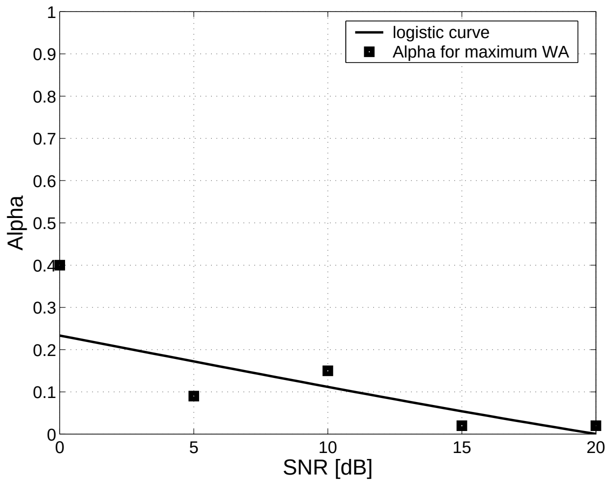

Fig. 5. Logistic curve for recognition using exhibition noise

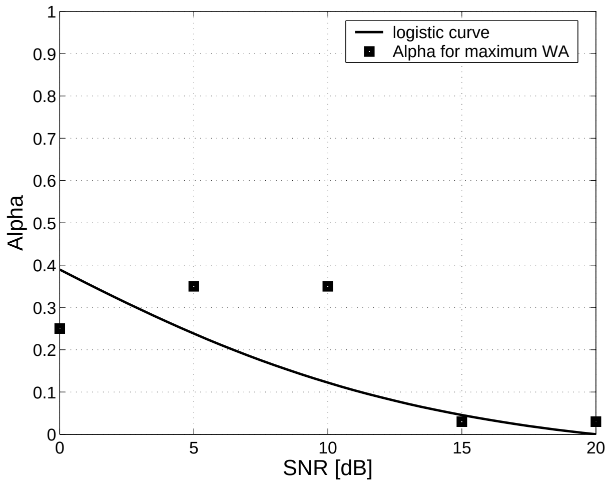

Fig. 6. Logistic curve for recognition using restaurant noise

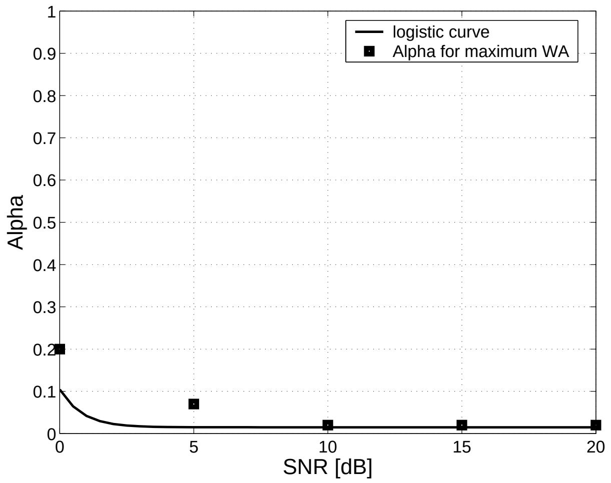

Fig. 7. Logistic curve for recognition using street noise

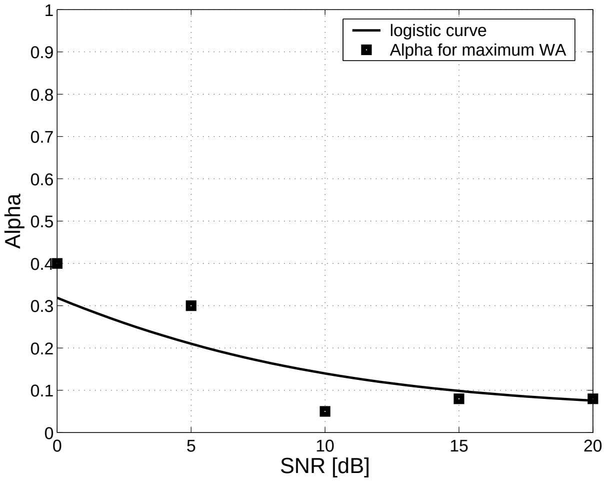

Fig. 8. Logistic curve for recognition using subway noise

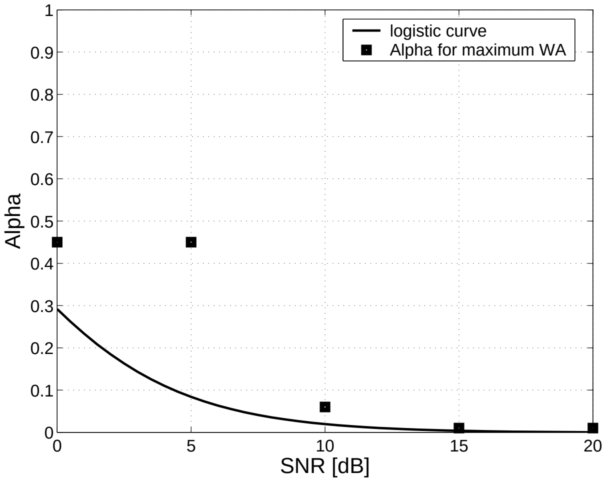

Fig. 9. Logistic curve for recognition using train noise

cost is reduced which is very important for real-life applications.

To validate the proposed logistic functions, recognition tests were performed for different preselected SNR levels using the canonical model adapted which their correspondent adaptation factors from respective curves. Experimental results showed that proposed algorithm led to a good adaptation coefficient value improving the system performance as can be observed in Tables X, XI and XII.

VI. Conclusions

Regarding the relationship between adaptation coefficient, noise type and level, besides strategy to determine its value, it is possible to conclude that the logistic function provides a good modeling of its behavior.

As a conclusion, we can state that the proposed strategy to choose an appropriate α value for a given type and noise level ensures an improvement of approximately 3% on system performance compared to the reference value (multi-style trained system), although it doesn't necessarily provide the best choice for this value.

TABLE X

Word Accuracy from logistic curve for SNR = 2 dB.

Noise	α from Logistic Curve	Reference WA (%)	WA using Logistic Curve (%)	ΔWA (%)
airport	0.4591	7.9	12.8	4.9
babble	0.1310	10.5	15.4	4.9
car	0.1147	12.5	15.9	3.4
exhibition	0.1865	2.4	6.9	4.5
restaurant	0.1401	6.1	7.5	1.4
street	0.0227	30.3	31.5	1.2
subway	0.2702	2.3	9.0	6.7
train	0.1844	12.3	15.6	3.3

TABLE XI

Word Accuracy from logistic curve for SNR = 7 dB.

Noise	α from Logistic Curve	Reference WA (%)	WA using Logistic Curve (%)	ΔWA (%)
airport	0.3333	39.8	47.3	7.5
babble	0.0366	48.7	50.3	1.6
car	0.0365	49.8	50.8	1.0
exhibition	0.0999	31.3	38.4	7.1
restaurant	0.0813	11.4	15.5	4.1
street	0.0149	65.9	65.9	0.0
subway	0.1777	33.9	40.8	6.9
train	0.0476	51.6	53.4	1.8

TABLE XII

Word Accuracy from logistic curve for SNR = 12 dB.

Noise	α from Logistic Curve	Reference WA (%)	WA using Logistic Curve (%)	ΔWA (%)
airport	0.2048	71.7	72.4	0.7
babble	0.0150	73.7	74.2	0.5
car	0.0271	73.9	73.9	0.0
exhibition	0.0450	69.1	69.6	0.5
restaurant	0.0505	53.2	54.9	1.7
street	0.0149	76.0	76.0	0.0
subway	0.1205	66.8	67.3	0.5
train	0.0104	74.6	74.6	0.0

References

[1] L. Lu, A. Ghoshal and S. Renals. Maximum a Posteriori Adaptation of Subspace Gaussian Mixture Models for Cross-Lingual Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4877-4880, March 2012.

[2] S. Furui. *50 years of progress in speech recognition technology: Where we are, and where we should go? From a poor dog to a super cat.* Keynote Presentation, ICASSP 2007.

[3] J. Rajnoha. Multi-Condition Trainin for unknown environment adaptation in robust ASR under real conditions. Acta Polytecnica, Vol. 49, No. 2-3, 2009.

[4] T. A. F. Valério. Treinamento multi-estilo e adaptação de modelos via MAP para reconhecimento de fala em ambientes ruidosos. Dissertação de Mestrado. Inatel. Santa Rita do Sapucaí, 2009.

[5] Huang, X., Acero, A. and Hon, H.W. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall. 2001.

[6] D. A. Reynolds, T. F. Quatieri and R. B. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. Digital Signal Processing, vol. 10, pp.19-41, 2000.

[7] H. G. Hirsch and D. Pearce. The Aurora experimental framework for the evaluation of speech recognition systems under noisy conditions. In Proc. ASR-2000, pp 181-188,September 2000.

[8] R. Lippman, E. Martin and D. Paul. Multi-style training for robust isolated-word speech recognition. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 12, pp. 705-708, 1987.

[9] L. Buera, E. Lleida, A. Miguel, A. Ortega and O. Saz. Cepstral Vector Normalization Based on Stereo Data for Robust Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 1098-1113, 2007.

[10] A. G. Chiovato. Avaliação da Relação entre Qualidade Perceptual da Fala e Taxa de Acerto de Sistemas de Reconhecimento de Fala em Ambientes Ruidosos. Dissertação de Mestrado. Inatel. Santa Rita do Sapucaí, 2005.

[11] C. A. Ynoguti and F. Violaro. Desenvolvimento de um Conjunto de Ferramentas para Pesquisa em Reconhecimento de Fala. Revista Telecomunicações, vol. 4(2), pp.36-43, 2001.

[12] A. Alcaim, J. A. Solewicz and J. A. Moraes, Frequência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro. Revista da Sociedade Brasileira de Telecomunicações, vol. 7(1), pp.23-41, 1992.

[13] H. Ney, The use of a one-stage dynamic programming algorithm for conected word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSSP-32(2), 1984.