IMPROVING THE ROBUSTNESS OF AUTOMATIC SPEECH RECOGNITION

Jean-Paul HATON

LORIA/INRIA
Université Henri-Poincaré, Nancy 1, France

Automatic speech recognition systems present performance that degrades dramatically in adverse situations, in the presence of noise or with different speakers. These degradations are due to the differences that occur between training and testing conditions. Speech recognition in adverse conditions has received increased attention during the last decade, since noise resistance has become one of the major bottlenecks for practical use of speech recognizers in real life. This paper reviews some of the methods proposed so far at the various stages of the recognition process in order to improve the robustness of the systems in real life conditions.

Key words: Automatic speech recognition, robustness, speech enhancement, parameterization, model adaptation, missing data.

1. INTRODUCTION

The performance of automatic speech recognition systems degrades dramatically in adverse situations, especially in the presence of noise or with different speakers. In particular, problems are created by the differences that may occur between training and testing conditions (noise level as measured by the signal-to-noise ratio (SNR), distance to the microphone and orientation, and also differences in speakers, etc.).

Speech recognition in adverse conditions has received increased attention for several years since noise resistance has become one of the major bottlenecks for practical use of speech recognizers in real life [27]. The methods proposed so far in order to improve the robustness of systems fall into different, non exclusive, categories: signal acquisition and parameterization, reference modeling, and adaptation techniques. After having briefly recalled the effects of noise on speech, we present some of the most common methods proposed in the literature to increase the robustness of systems.

2. EFFECT OF NOISE ON SPEECH

The various kinds of noise cause substantial alterations to the speech signal. The main sources of speech variation can be classified into three main categories:

- *addition of ambient noise:* it is usually admitted that a recorded speech signal is the sum of the speech produced by a speaker and the ambient noise. This noise is usually a colored noise and its structure can vary significantly, according to the source: office machinery (typewriters, workstations, etc.), human conversations (babble noise), car (coming from engine, wind, tires, road, etc.), plane cockpit, industrial plant, etc;
- *distortion of the signal:* the speech signal undergoes various distortions that may affect its frequency structure and phase in a usually non-linear way. Such distortions result from the convolution of the speech signal with a particular system. They can for instance be produced by room reverberation. Microphone transduction can also distort the speech spectrum in a way specific to each type of microphone and mounting position. Therefore, the use of different microphones for training and testing can lead to significant spectrum mismatch, and causes important discrepancies in recognition. Finally, in telephony applications, the transmission channel can also cause speech distortions, mainly through a frequency-dependent attenuation. The resulting distortions in the speech spectrum, or spectral tilt, are a major cause of performance degradation in automatic speech recognition. Some of the methods proposed so far are able to carry out simultaneously a compensation of noise and spectral

tilt. Such a joint compensation has been shown as more effective than a combination of independent compensators;

• *variations in articulation:* a speaker can be affected in his speaking manner by different factors like stress, emotion, physiological state, etc. But the most important factor is perhaps the influence of a noisy environment. When speakers speak under heavy noise and/or stress conditions, they dramatically change their utterance in terms of formant frequencies, pitch, sound duration, etc. This Lombard effect has a strong influence on the performances of a speech recognizer, even if speakers can be trained to some extent to avoid to some extent Lombard speech in noisy environments.

3. OVERVIEW OF PROBLEMS AND SOLUTIONS

Most present speech recognition systems rely on a statistical framework. The basic idea is to compute the conditional probability P (W/O) of recognizing a sequence of words W for an acoustic input signal O, thanks to Bayes formula. This computation involves two types of models:

- acoustic models, usually under the form of Hidden Markov Models (HMM),
- language models, usually under the form of n-grams models.

This paper is concerned with the increase in robustness of acoustic models. Increasing the robustness of language models, especially with adaptation techniques is also an important issue, but that lies outside the scope of this paper (see for instance [6] for a review of the problem).

The difficulties encountered in the design of a speech recognition system come from the various effects of noise on speech described in section 2. They result in two major phenomena:

• the degradation of the performance of an automatic speech recognition system if the learning and testing conditions are different. For example, it has been found a degradation of an order of magnitude in the error rate for an isolated word recognizer when trained with clean speech and tested with noisy utterances at a signal-to-noise ratio of 18 dB. This requires the improvement of the overall robustness of speech recognition, because the simple solution of having the same conditions for training and testing is seldom affordable and usually unrealistic;

• the modification of sounds pronounced in a noisy environment (the already mentioned Lombard effect). This effect is highly dependent upon the speaker, the context and the level of noise, and is thus very difficult to quantify and to model. Its main effects are increases of fundamental frequency and of first and second formants, changes in the spectral tilt, increases in formant amplitude, variations in phoneme and word durations, etc. Several experiments have confirmed the large degradation of speech recognition performance due to the Lombard effect.

In brief, the mismatch between training and testing patterns is the major problem to deal with in adverse, real life conditions of speech recognition. A large number of methods have been proposed to cope with these problems, even though none is totally satisfactory. These methods are used at some steps in the basic sequence of speech recognition processing. Of course, the different methods are not exclusive, and can be combined in order to obtain satisfactory performance.

4. PREPROCESSING AND PARAMETERIZATION

4.1. Position of the problem

The overall performance of a speech recognizer highly depends upon the quality and robustness of the acoustic and sometimes phonetic features extracted from the speech wave as a front end of recognition algorithms. Therefore a major effort has been devoted to this problem.

All existing methods are not easily usable in adverse conditions. For instance, basic LPC analysis is very sensitive to noise. New models, or adapted versions of classical models, are thus necessary.

Recently, an international effort of standardization within the Aurora ETSI group has resulted in the selection of a new standard for a robust front-end. The system that was selected features an optimal combination of noise reduction, SNR dependent waveform processing, cepstrum calculation, blind equalization, and voice activity detection [33].

4.2. Voice activity detection

Substantial improvements in recognition accuracy can be obtained by introducing an accurate Voice Activity Detector, robust to noisy conditions. Linear estimation and logistic regression [44] are among the most popular methods for computing a probability of the presence of speech in a signal. Other proposed approaches are based on Higher Order Statistics (HOS), log likelihood ratio, estimation of the *a priori* speech absence probability controlled by the minima values of the smoothed power spectrum of the noisy signal, Laplacian Gaussian model.

An Artificial Neural Network (ANN) has also been proposed for computing a speech/no-speech probability for each speech frame [5], scaled and used as an additional observation for acoustic Hidden Markov Models. Another ANN has also been proposed for computing a Noise Presence Probability in absence of speech with two major differences [18]. This probability is used in a non-linear gain function that depends on the estimated strength of the non-speech state of the acoustic model. The function is used in a hybrid ANN/HMM ASR system to modify the non-speech posterior probability computed by the ANN of the hybrid system. As a consequence, the posterior probabilities of all the other phone units are simply normalized, without affecting their relative values computed with a model trained with a discriminative learning process. In this way, the amount of speech/non-speech confusion is significantly reduced.

4.3. Speech Enhancement

As a first step in the recognition process, speech enhancement techniques tend to suppress the noise that corrupts the speech signal. Besides these methods using several microphones, many different types of speech enhancement systems using a single microphone have been proposed and tested. All these systems are based on techniques designed to recover the clean speech signal by enhancing the signal-to-noise ratio. The performance depends upon the type of noise that corrupts speech and the information they require about noise. It should be noted at this point that the increase of SNR will improve the quality of the speech signal without always improving its intelligibility. Therefore, as far as speech recognition is concerned, a trade-off has to be found between SNR improvement and recognition accuracy. Moreover, speech enhancement techniques have initially mainly dealt with the improvement of speech quality and intelligibility for human listeners. Even though the problem is not totally similar, these techniques will also improve automatic recognition.

Several types of methods are used for speech enhancement:

- *noise subtraction:* this a very common method based on the assumption that noise and speech are uncorrelated and additive [7]. In the spectral subtraction approach, the power spectrum of cleaned speech is obtained by subtracting the noise power spectrum from the spectrum of noisy speech. The noise spectrum is estimated during pause intervals by averaging short-term power spectra over successive frames. The method assumes that the noise varies slowly so that the noise estimation obtained during a pause can be used for suppression. Obtaining a good estimate of the noise spectrum is obviously the most difficult part of the method.

This process is quasi-linear: the only non-linearity is introduced by the specific solutions to the avoidance of negative spectral magnitudes in the subtraction operation (e.g. thresholding). There are several biases in the process: thresholding, approximation on signal phase (the phase of noisy signal is used for reconstructing the spectrum of cleaned signal), non-stationarity of real noise, etc.). These biases result in the introduction of a "musical" noise in the cleaned signal due to the presence of spurious peaks in the spectrum. A solution to this problem is to design a non-linear spectral subtraction which basically consists in overestimating the noise spectrum, either in a uniform way, or else based on the perceptual evidence that the ear is more sensitive to the peaks of a power spectrum than to the valleys and that noise in the frequency regions of the valleys contributes the most to perceptual distortions. This latter solution has significantly improved the recognition performances compared to normal subtraction [32]. Non-linear subtraction has given very good results in noisy word recognition, especially when combined with robust recognition techniques;

- *filtering:* traditional adaptive filtering techniques like Wiener or Kalman filtering have been used for speech enhancement, but more for speech transmission than for recognition purposes. As for the noise subtraction techniques, the most difficult aspect is the proper estimation of noise characteristics from observations. The Wiener filter provides an optimal solution to the adaptive filtering problem in the sense of least mean square error. It necessitates the estimation of some parameters of the noise. Unless the noise is stationary and perfectly known that must usually be done iteratively. A recursive optimal

estimation can be obtained with a Kalman filter. In this method an AR speech model is estimated by an autocorrelation method at each speech frame s_n. This speech model together with the noise model are used by a Kalman filter in order to extract from s_n a better estimate of the enhanced signal $\hat{s}_n$. This method results in a substantial improvement in the SNR but, at the same time, it introduces several distortions in the speech signal that deteriorate the recognition rate and lead to performances lower than those obtained with non-linear spectral subtraction;

• *space mapping:* speech enhancement can be viewed as the process of transforming noisy speech into clean speech by some kind of mapping. For instance, spectral mapping has been implemented by a set of rules obtained by vector quantization techniques. A statistical mapping of spectra has also been done. This technique is based on the extraction of noise-resistant features from speech, and can thus be considered as noise-independent to some extent. Another method based on the use of multiple linear regression techniques applied to MFCC vectors has been reported to give better results than linear spectral subtraction for word recognition in a car. The idea can be generalized to arbitrarily complex space transformations thanks to connectionist neural networks. Even simple models such as multi-layer perceptrons have been trained on learning samples to produce a mapping of noisy signals to noise-free speech that has been tested successfully in an auditory preference test with human listeners. They have also been used for learning to differentiate between similar patterns such as plosive consonants in the recognition of noisy spelled letters, and for normalizing speech data in order to adapt a recognizer to variations in telephone line conditions. Some improvements have been brought to the method based on the fact that the separation between speech and noise is easier to carry out at the output of a hidden layer of the network rather than in the initial physical space.

4.4. Improvement of Speech Analysis Methods

An important category of noisy speech processing techniques is concerned with the design of robust front-ends that produce noise-resistant acoustic features. Such methods usually do not make any assumptions about the characteristics of noise. Some of them are described below.

• *non-parametric representations:* Mel Frequency Cepstrum Coefficients (MFCC) are to some extent resistant to noise, and certainly more than conventional LPC analysis. Their performances are significantly improved by adding dynamic features (Δ MFCC), i.e. the temporal slopes obtained by regression on the MFCC coefficients [13]. More generally, the use of dynamic and acceleration (second derivative) MFCC and energy features makes it possible to enhance the recognition performances for noisy and/or Lombard speech. It is worth noticing that time derivatives are high-pass functions. These techniques are therefore related to the subband or cepstral domain filtering methods such as RASTA (cf. below), subband high-pass filtering, or spectral normalization.

A general framework for removing distortions from noisy speech cepstra, called SPLICE [11], has also been designed. SPLICE takes noisy acoustic observations y, and produces clean estimates x. The relationship between x and y is modelled as a constrained Gaussian mixture model (GMM). An auxiliary variable m is introduced to index the hidden state of the GMM. The SPLICE transform $f_\mu(y)$ is the MMSE estimate of x, given y and the model parameters μ. There have been several investigations in producing discriminatively trained linear transformations of the feature space, such as MCMIP [36], MCELR [19] and MPE-HLDA [46]. These methods achieve relatively modest improvements over LDA and HDA. A natural extension is to discriminatively train a non-linear transformation of the feature space, instead linear ones. This is the approach taken by MCE-SPLICE [45], fMPE [37], and MMI-SPLICE [12]. This eliminates the necessity of providing observable clean features x. While ML-SPLICE learns a transformation to a fixed oracle feature space, MMI-SPLICE is only concerned with improving the accuracy of the end-to-end system.

A mechanism of feature compensation has been proposed for enhancing the robustness of MFCC. The basic idea of compensation is to subtract features that are dependent on a particular speaker or environment from observed features. In cepstrum mean normalization (CMN), the long-term average of the cepstrum is subtracted from each speech frame. This helps eliminate changes due to differences among speakers as well as to environment and channel changes, since these changes are slower than the changes in the features of speech itself [2].

Data analysis techniques have been used in the IMELDA system [23], [24] in order to obtain a robust representation of noisy speech (but also of clean speech, as further experiments have demonstrated). IMELDA carries out a linear transformation based on discriminant analysis with minimization of within-

class differences and maximization of inter-class differences. The result is a low dimensionality representation space in which recognition algorithms perform well. An advantage of the method is that recognition is also computationally inexpensive. IMELDA coefficients are computed from an initial parameter space which varies according to the version of the system: outputs of a large filter bank, static and dynamic outputs of the filter bank, outputs of a simulated auditory model based on Seneff's model. This latter representation has given good performances for both clean and noisy speech recognition. The combination of IMELDA with a non-linear spectral subtraction method has also been shown as giving improved performances in the recognition of speech in a car;

- *parametric representations:* several improvements of LPC analysis in the presence of noise have been tested with some success. Some of these methods are based upon an alternative solution to the speech deconvolution problem. The classical solution consists in identifying the impulse response of the vocal tract by AR or ARMA modeling. The other solution is to map the time signal space into a linear structure by using a homomorphic transformation that corresponds to a filtering in the cepstral domain. The logarithm homomorphic deconvolution can be further generalized to a spectral root deconvolution. This root scheme differs from the original homomorphic deconvolution scheme by changing the logarithmic and exponential functions respectively by $(.)^{ß}$ and $(.)^{1/ß}$ with $-1 < ß < +1$, $ß \neq 0$. This scheme has been proven as significantly less affected by noise than the log scheme.

The short-time modified coherence, SMC [34] representation of speech computes coefficients that are LPC coefficients obtained by taking the characteristics of the autocorrelation domain into account. It can be best described as the estimate of an AR model in the autocorrelation domain, and it exploits the fact that an autocorrelation sequence is more resistant to noise than the original signal. SMC analysis made it possible to increase the SNR by 10 to 12 when it was initially 0 to 20 dB. It is worth noticing that the amount of improvement is essentially data dependent.

SMC has some drawbacks, especially a stronger interaction between F0 and formant frequencies due to the fact that computation is done in the spectral power density domain. That can be overcome by introducing in the algorithm a square root calculation instead of the absolute value of FFT that makes some kind of spectral compression. SMCC coefficients can also be obtained by applying a cepstrum calculation to the SMC coefficients.

PLP (Perceptual Linear Prediction) [20] differs from LPC by the use of three concepts derived from the study of human hearing, i.e. critical band spectral resolution, pre-emphasis with an equal-loudness curve, and spectral compression according to a intensity-loudness power law. A comparative study has shown the superiority of PLP over LPC for noisy speech recognition, especially in conjunction with a liftered cepstral distance (cf. section 6) [26].

The RASTA (Relative Spectral) [21] operates both in time and frequency. The method consists in operating in the log power spectral domain. That makes it possible to remove, or at least efficiently reduce, by filtering techniques, slow-varying communication noise, which is additive in the log domain. On the other hand, noise which is additive in the time domain will not be removed, and could possibly be exaggerated by the log operation. A proposed solution consists in designing a filtering function that is linear for low values in the auditory spectrum and approximately logarithmic for large values with a threshold related to the SNR value. This method has been tested on small vocabulary, isolated word recognition. A combination of PLP and RASTA analysis has also been reported as effective for speaker verification with degraded speech [22].

RASTA is a good way for capturing the temporal properties of speech. But the speech signal also exhibits combined spectro-temporal modulations due to various factors: intonation, coarticulation, etc. A psychoacoustical model, implemented with the help of two-dimensional Gabor functions, has been proposed in order to capture such modulations. This model has provided a consistent performance improvement [28].

5. MISSING DATA APPROACHES

Non-stationary noise is particularly difficult to handle in ASR systems. Speech recognition with missing data was proposed to deal with such conditions. The idea is to take advantage of the high redundancy of the speech signal. Usually, noise affects only localized zones of the time-frequency plane, so that there remains sufficient information in the clean regions of the spectrum to recognize speech. The least noise-corrupted features are labeled as reliable, whereas the noisiest ones, associated with missing data are discarded.

There are two different methods to handle missing data during the recognition process. The first method, called *data imputation*, consists in replacing all missing values by their estimation in order to provide the recognizer with a completely observed signal. Several methods have been tried to infer the missing values [39]. In the second method, called *data marginalization*, the recognition engine is modified in such a way that it could handle partial feature vectors. Several algorithms for data marginalization have also been proposed [4], [35].

This missing data approach necessitates classifying every input feature as reliable or missing, hence defining a so-called missing data mask. The overall recognition performance is highly dependent upon the quality of these masks. Several mask estimation methods have been proposed so far [9], [40]. Among them, Bayesian mask estimation consists in training statistical models for reliable and unreliable features and in estimating the most probable mask according to the noisy observed feature. An original Bayesian mask model has recently been proposed that takes into account two major properties [9]:

- a set of features that contains information from spectrographic neighbors is assigned to each coefficient;
- dependencies across frequency and time between mask values are introduced during the mask estimation process. The proposed solution is global, which means that mask values that are far from the coefficient of interest may influence the decision of the mask estimator.

This proposed solution has been tested on Aurora 2 and 4 databases. Experimental results show that the dependencies between mask values improve recognition accuracy.

6. MODEL ADAPTATION

Since a major cause of performance degradation is the discrepancy between training and testing conditions, it seems interesting to transform the parameters of recognition models in order to adapt them to new conditions. Such adaptation techniques have received considerable interest during the past few years, as well for speaker adaptation [14] as for environment adaptation, either in a supervised or unsupervised mode.

One of the first of this category is the Parallel Model Combination (PMC) scheme [15]. This method applies to stochastic models like HMMs or trajectory models. It consists of choosing a representation space that makes it possible to obtain an adapted noisy model by simple combination of a clean speech model and of a noise model (the noise being supposed stationary). Combination methods need to have sufficient noise data to train the model. They suffer from the problems of unknown noise types or Signal-to-Noise Ratios. Some solutions have been proposed to solve these problems [25].

Regression methods have also been extensively used to adapt the parameters of a model. Among the most popular is the linear regression by maximum likelihood (*Maximum Likelihood Linear Regression, MLLR)*. In this method, the means of the statistical model, and possibly the variances, are estimated as linear combinations of the original means and of a bias in order to linearly transform the probability density functions of acoustic models. These estimations are based on a set of adaptation data with a maximum likelihood (ML) criterion [31]. For the performance improvement obtained by unsupervised MLLR on Aurora 4 task, see [3].

Another class of methods consists of a *maximum a posteriori* Bayesian estimation [17]. In that case, the probability distribution function of the model parameters is chosen *a priori*. Parameters are then estimated with a *Maximum a Posteriori* (MAP) criterion, instead of a classical maximum likelihood. This method can be viewed as an extension of the training phase limited to the adaptation data and some prior information about the estimation parameters. This information is particularly useful in case of sparse adaptation data.

Several approaches have been proposed in order to cumulate the advantages of MLLR and MAP techniques [1], in particular Structural Maximum a Posteriori, SMAP [42], Maximum a Posteriori Linear Regression, MAPLR [10], and Structural Maximum a Posteriori Linear Regression, SMAPLR [43]. All these methods exploit the hierarchical organization of acoustic models and the hierarchical principle of parameter estimation under the MAP criterion in order to improve the estimation. Available adaptation is more efficiently used, and, moreover, the models for which no adaptation data are present can nevertheless be adapted. Such approaches are particularly useful when a rapid adaptation is necessary, for instance in telephone-based dialogue systems.

The adaptation to new data can also be carried out by a first order Jacobian approach in which initial acoustic models under an initial condition A are to be brought to a target condition B, assuming that A and B are not too different. This method has been shown to be effective for adaptation to a given noisy condition [41], [8].

Another approach to speaker adaptation is based on clustering techniques. The basic idea consists of clustering the speakers of the training set according to acoustic similarity in order to capture intra-speaker inter-class correlations [14]. The Speaker Cluster Weighting algorithm, SCW [16] extends this idea by allowing the speaker cluster models that have been learned from the training data to be adaptively weighted to match the current speaker. A combination of SCW with a clustering method for dealing with speakers for whom a small amount of data is available has also been proposed [38]. The SCW approach is similar in principle to eigenvoices [30] in the sense that this latter also explicitly utilizes the set of training data to constrain the adaptation.

7. CONCLUSION

A large variety of methods have been proposed so far in order to increase the robustness of automatic speech recognition systems. This problem is very difficult and diverse; it constitutes a major bottleneck for the practical use of speech recognizers in real conditions. In this paper, we have briefly presented some of the most promising approaches proposed so far.

Despite significant results, several questions are still open. As a matter of fact, it can be said that the problem of robust, environment-independent or environment-adaptive speech recognition is still in its infancy. A major issue is the dependability of present methods to noise characteristics or to speakers. Another important point is the need for methods capable of dealing with non-stationary noise conditions, such as door slams, telephone rings or other transitory sounds in an online, unsupervised way.

A large effort has still to be done in order to comparatively evaluate the different existing methods, and to improve them. This necessitates an important, and on-going, effort of speech databases recording and assessment method design.

REFERENCES

1. AFIFY M., GONG Y., and HATON J. P., “A Unified Maximum Likelihood Approach to Acoustic Mismatch Compensation: Application to Noisy Lombard Speech Recognition”, *Proc. ICASSP*, Munich, 839-842, 1997.
2. ATAL B., “Effectiveness of Linear prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification”, *J. Acoust. Soc. Amer.* **55**, 1304-1312, 1974.
3. AU YEUNG S.K., and SIU M. H., “Improved Performance of Aurora 4 using HTK and Unsupervised MLLR Adaptation”, *Proc. ICSLP,* Jeju, 2004.
4. BARKER J. *et al.*, “Soft decision in missing data techniques for robust speech recognition”, *Proc. ICSLP*, Beijing, 2000.
5. BEAUFAYS F. *et al.*, “Using Speech/Non-speech detection to bias recognition search on noisy data”, *Proc. ICASSP*, Hong Kong, 424-427, 2003.
6. BELLEGARDA J. R., “An Overview of Language Model Adaptation”, *Proc. ISCA Automatic Speech Recognition Workshop*, 2001.
7. BOLL S. F., “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, *IEEE Trans. ASSP* 27, 113-120, 1979.
8. CERISARA C. *et al.*, “Transformation of Jacobian Matrices for Noisy Speech Recognition”, *Proc. ICASSP*, Salt Lake City, 2001.
9. CERISARA C., DEMANGE S., and HATON J. P., “On noise masking for automatic missing data speech recognition: A survey and discussion”, *Computer Speech and Language*, in press, 2007.
10. CHESTA O. S. T., “Maximum a Posteriori Linear Regression for Hidden Markov Models Adaptation”, *Proc. Eurospeech*, 211-214, 1999.
11. DROPPO J. *et al.*, “Evaluation of SPLICE on the Aurora 2 and 3 Tasks”, *Proc. ICSLP*, Denver, 29-32, 2002.
12. DROPPO J., and ACERO A., “Maximum Mutual Information SPLICE Transform for Seen and Unseen Conditions”, *Proc. Interspeech*, 2005.
13. FURUI S., “Speaker Independent isolated Word Recognition Using Dynamic Features of Speech Spectrum”, *IEEE Trans. ASSP* 34, 52-59, 1986.
14. FURUI S., “Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering”, *Proc. ICASSP*, Glasgow, 286-289, 1989.
15. GALES M., and YOUNG S., “HMM Recognition in Noise Using Parallel Model Combination”, *Proc. Eurospeech*, 837-840, 1993.
16. GAO Y. *et al.*, “Speaker Adaptation Based on Pre-clustering Training Speakers”, *Proc. ICSLP*, Denver, 61-64, 2002.
17. GAUVAIN J. L., and LEE C. H., “Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, *IEEE Trans. Speech and Audio Processing* 2, 291-298, 1994.
18. GEMELLO R. *et al.*, “Non-linear estimation of voice activity to improve automatic recognition of noisy speech”, *Proc. Interspeech*, Lisbon, 2005.
19. HE X., and CHOU, W., “Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs,” in *Proc. ICASSP*, Hong Kong, 556–561, 2003.

20. HERMANSKY H., "An Efficient Speaker-independent Automatic Speech Recognition by Simulation of Some Properties of Human Auditory Perception", *Proc. ICASSP*, Dallas, 1159-1162, 1987.
21. HERMANSKY H. *et al*., "Compensation for the Effect of Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP)", *Proc. Eurospeech*, Genève, 1367-1370, 1991.
22. HERMANSKY H., and MORGAN N., "RASTA Processing of Speech", *IEEE Trans. Speech and Audio Processing* 2, 578-589 (1994).
23. HUNT M. J., and LEFEBVRE C., "A Comparison of Several Acoustic Representations for Speech with Degraded and Undegraded Speech", *Proc. ICASSP*, Glasgow, 1989.
24. HUNT M. J., RICHARDSON S. M., BATEMAN D. C., and PIAU A., "An Investigation of PLP and IMELDA Acoustic Representations and of Their Potential for Combination", *Proc. ICASSP*, Toronto, 881-884, 1991.
25. IDA M., and NAKAMURA S., "HMM Composition-Based Rapid Model Adaptation Using a Priori Noise GMM Adaptation Evaluation on Aurora2 Corpus", *Proc. ICSLP*, Denver, 437-440, 2002.
26. JUNQUA J. C., and WAKITA H., "A Comparative Study of Cepstral Lifters and Distance Measures for All Pole Models of Speech in Noise", *Proc. ICASSP*, Glasgow, 476-479, 1989.
27. JUNQUA J. C., and HATON J. P., *Robustness in Automatic Speech Recognition: Fundamentals and Applications,* Kluwer Academic, 1996.
28. KLEINSCHMIDT M., and GELBART D., "Improving Word Accuracy with Gabor Feature Extraction", *Proc. ICSLP*, Denver, 25-28, 2002.
29. KOSAKA T., and SAGAYAMA S., "Tree-Structured Speaker Clustering for Fats Speaker Adaptation", *Proc. ICASSP*, Adelaide, 245-248, 1994.
30. KUHN R. *et al.*, "Rapid Speaker Adaptation in Eigenvoice Space", *IEEE Trans. On Speech and Audio Processing* 8, 695-707 (2000).
31. LEGGETTER C. J., and WOODLAND P. C., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language* 9, 171-186 (1995).
32. LOCKWOOD P. *et al.*, "Non-linear Spectral Subtraction (NSS) and Hidden Markov Models for Robust Speech Recognition in Car Noise Environments", *Proc. ICASSP*, San Francisco, 265-268, 1992.
33. MACHO D. *et al.*, "Evaluation of a Noise-Robust Front-End on Aurora Databases", *Proc. ICSLP*, Denver, 17-20, 2002.
34. MANSOUR D., and JUANG B. H., "The Short-time Modified Coherence Representation and its Application for Noisy Speech Recognition", *Proc. ICASSP*, New York, 525-528, 1988.
35. MORRIS A., "Data utility modeling for mismatch reduction", *Proc. Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis,* 2001.
36. OMAR M., and HASEGAWA-JOHNSON M., "Maximum conditional mutual information projection for speech recognition", Proc. *Eurospeech*, Genève, pp. 505–509, 2003.
37. POVEY D. *et al.*, "FMPE: Discriminatively trained features for speech recognition", *Proc. ICASSP*, Philadelphia, 961–964, 2005.
38. PUSATERI E. J., and HAZEN T. J., "Rapid Speaker Adaptation Using Speaker Clustering", *Proc. ICSLP*, Denver, 61-64, 2002
39. RAJ, B., *Reconstruction of incomplete spectrograms for robust speech recognition*, PhD thesis, Carnegie-Mellon University, 2000.
40. RAJ B., and STERN R. M., "Missing features approaches in speech recognition", *IEEE Magazine*, 2005.
41. SAGAYAMA S. *et al.*, "Jacobian Approach to Fast Acoustic Model Adaptation", *Proc.ICASSP*, Munich, 835-838, 1997.
42. SHINODA K., and LEE C. H., "A Structural Bayes Approach to Speaker Adaptation", *IEEE Trans. On Speech and Audio Processing* 9, 276-287 (2001).
43. SIOHAN O., and LEE C. H., "Structural Maximum a Posteriori Linear Regression for Fast HMM Adaptation", *Proc. ISCA Automatic Speech Recognition Workshop*, 120-127, 2000.
44. SURENDRAN A. C., SUKITTANON S., and PLATT J. C., "Logistic Discriminative Speech Detectors Using Posterior SNR", *Proc. ICASSP*, Montreal, 2004.
45. WU J., and HUO Q., "An environment compensated minimum classification error training approach and its evaluation on Aurora2 database," *Proc. ICSLP*, Denver, 453–457, 2002.
46. ZHANG B. , and MATSOUKAS S., "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," in *Proc. ICASSP*, Philadelphia, 925–929, 2005.