Statistical Techniques for Robust ASR: Review and Perspectives

Jerome R. Bellegarda

Advanced Technology Group, Apple Computer,
Cupertino, California 95014, USA

jerome @ apple.com; +1 (408) 974-7647

ABSTRACT

Speech recognition performance degrades significantly when a mismatch occurs between training and operating conditions. To reduce this mismatch, it is often necessary to characterize the mapping between the two environments. A number of statistical approaches have been developed for this purpose. They can be classified as either predictive or adaptive, depending on what information is available regarding the operating environment. This paper reviews a selected subset from both categories, and discusses possible future directions of improvement.

1 INTRODUCTION

Speech recognition performance is known to degrade significantly when a mismatch occurs between training and operating conditions; see, e.g., [1]-[4]. This mismatch can result from differences in: (i) transmission characteristics, such as the type of microphone selected and the available bandwidth [2]; (ii) background acoustic environment, comprising external noise sources as well as reflection and reverberation effects [3]; and (iii) articulatory phenomena resulting from a change in speaker or speaking style, including those due to the Lombard effect [4]. Fig. 1 gives a simplified depiction of how the speech signal may be affected.

Three broad stategies have been followed to reduce this mismatch [3]. First, search for features and metrics that are inherently robust, so the mismatch is irrelevant. Second, retrieve the original speech from the corrupted speech, so as to operate in the original training conditions (*speech enhancement*). Third, appropriately "corrupt" the original training parameters, so as to perform recognition in the current operating environment (*noise compensation*). In the latter two approaches, it is necessary to characterize the mapping between the two environments. We concentrate here on some of the techniques that have been developed to model this mapping, estimate the relevant parameters, and make adequate adjustments between training and recognition conditions. Compared to other surveys (e.g., [1]-[4]), this paper is focused on statistical approaches, and follows a different classification.

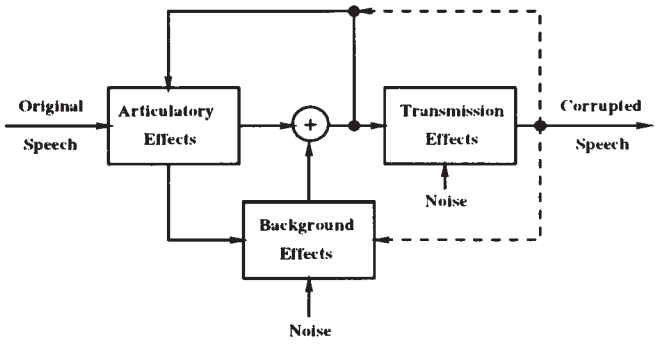

Fig. 1. Sources of Variability in the Speech Signal.

First, since our main goal is to characterize the mapping between training and operating environments, the distinction between speech enhancement and noise compensation is not critical, as it depends only on the direction of the transformation. Specifically, speech enhancement techniques transform the observed speech parameters (in this case, features) into the environment seen during training. In contrast, noise compensation techniques transform the original speech parameters (in this case, either features or models) into the current operating environment. By and large, we will not distinguish between the two.

Second, statistical approaches developed for one type of mismatch (e.g., speaker) can often be applied to another (e.g., background noise). For example, although "speaker" and "noise" issues have long been approached separately, they have converged in recent years. (In Fig. 1, speaker variability can be implicitly addressed as a subset of articulatory effects.) As a result, we will attempt to be indifferent to the various types of mismatch that can occur.

Finally, because sources of variability affect the speech signal in many ways (cf. Fig. 1), they are typically addressed at various levels of parameterization, such as the (linear or log) spectral, (LPC or mel) cepstral, or probabilistic model (HMM) domain. On the other hand, several approaches have been successfully applied to multiple feature and model spaces [3]. In the following we will tend to abstract out the parameterization selected,

and refer the reader to the original reference(s) for the details of implementation.

Thus, we have chosen to classify techniques based on the amount of a *priori* information available about the operating environment. Approaches tailored to a specific environment will be called *predictive*. Approaches that can adapt to any environment will be termed *adaptive*. Note that techniques classified to these two categories may overlap, as different motivations and objectives sometimes lead to similar solutions.

The paper is organized as follows. In the next section we review some of the predictive approaches that have been proposed over the past few years, such as feature mapping, probabilistic noise masking, and parallel model combination techniques. Then, Section 3 reviews some of the adaptive approaches, such as simple bias removal, stochastic matching, linear regression, and non-linear transformation techniques. Finally, in Section 4 we offer some perspectives and discuss possible directions of improvement.

2 PREDICTIVE APPROACHES

In this section we assume that specific information is available regarding the operating environment considered. In practice, this information takes the form of stereo recordings (whether actual or artificial), or, alternatively, a *priori* restrictions on the noise present.

2.1 Feature Mapping

When stereo recordings are available, it is straightforward to relate a speech event observed in the training environment with its counterpart observed in the operating environment. This can be used to derive a mapping between the respective feature vectors, without the need for explicit speech and noise models, nor the way they are combined.

In probabilistic optimum filtering [5], for example, the mapping is given by a piecewise linear transformation, estimated by quantizing the feature space into a set of distinct regions and computing a set of filters optimum in the mean square error sense.

An alternative technique was described in [6], where training and operating observations were each modeled as a set of random sources, with cross-correlations between them. A solution to the joint probability density function of the training observation, operating observation, and parameters of the respective random sources was proposed using the expectation-maximization algorithm. This probability density function was then used to find the minimum mean square estimate of the mapped feature vector.

As a remark, let us also mention that arbitrarily complex transformations can be achieved using neural networks. Examples include [7] and [8].

2.2 Probabilistic Noise Masking

When speech is observed in the presence of noise, it is inevitable that some useful information may be completely lost due to the noise corruption process. The traditional noise masking approach postulates that no information about speech can be derived from observations that fall below a particular noise level, assumed to be known with certainty. The integrated signal-background model described in [9] generalizes this approach by allowing uncertainty about the background noise.

This model provides a probabilistic framework to derive some level of information from the corrupted observations. Both signal and noise are assumed to be Gaussian mixture processes, and the feature parameters are estimated via maximum likelihood. This approach can therefore be viewed as an instance of speech and noise decomposition, applied at the feature level [10].

2.3 Parallel Model Combination

Applying speech and noise decomposition at the model level naturally leads to parallel model combination (PMC). With this technique, noisy speech signals are modeled using composed HMMs derived from individual speech and noise HMMs [11]. Thus, extending standard search algorithms to the composed HMMs allows simultaneous recognition of both signal and noise.

Because model decomposition provides a framework for incorporating independent concurrent processes, this approach can potentially deal with any non-stationary interfering signal. For example, multi-state noise HMMs can be specified to reflect fast changing and impulsive statistical characteristics. In addition, the acquisition of noisy speech is not necessary and when noise changes, only the specification of the noise model topologies and parameters needs to be changed [12]. However, as the noise models become more complex, the technique becomes computationally expensive.

3 ADAPTIVE APPROACHES

In this section we do not make explicit predictions of the noise parameters, nor do we require stereo recordings. We only assume that a "small" amount of adaptation data is available for adaptation purposes.

3.1 Simple Bias Removal

In the feature space, simple bias removal (SBR) can be viewed as a way to extend and improve upon cepstral mean normalization. In [13], for example, an iterative procedure was proposed for estimating a cepstral bias using maximum likelihood. The method was integrated into a discrete density HMM system, and can be applied to the spectral domain as well. Other developments along the same lines include the various versions of codeword dependent cepstral normalization (CDCN). To jointly compensate for both additive and convolutional noise, the CDCN approach consists of two phases, one to estimate the environmental parameters, and the other to derive a minimum mean square estimate of the clean speech, which is usually represented by a Gaussian mixture model [14]. In its standard form, CDCN does not require any *a priori* information about the environ-

ment. However, the computational load often leads to some approximations, which sometimes effectively assume availability of stereo data.

Similar approaches have been used in the context of speaker adaptation. In [15], a fixed bias was estimated to transform each individual speaker to a reference speaker, and the estimated bias was then subtracted to every frame of speech. In the model space, the transfer vector field approach [16] follows a somewhat analogous strategy. This technique estimates the differences between the means of associated HMM Gaussian distributions. This is particularly effective for small adaptation sets since it allows for interpolation and smoothing of the transfer vectors so obtained. Note that the use of maximum a posteriori estimation [17] allows the inclusion of prior information from an initial model.

3.2 Stochastic Matching

Compared to simple bias removal, stochastic matching models the mismatch as a random bias, which requires an additional variance estimation [18]. As before, this approach can be used either in the feature space or in the model space. To enable unsupervised adaptation, maximum likelihood estimation is applied in an iterative fashion. In addition, the method can accommodate separate speech and silence biases [19].

While this technique was originally developed for telephone speech, where convolutional noise is dominant [19], it has also been proven effective against additive noise. In [20], for example, it is shown that multivariate Gaussian-based cepstral normalization ("blind RATZ") performs well despite the fact that the noisy observations do not follow a Gaussian distribution [21].

3.3 Linear Regression

Statistical linear regression approaches are closely linked to the stochastic matching framework just described. A currently popular form is maximum likelihood linear regression (MLLR), in which the corrupted means are expressed as a (linear) function of both the original means and a bias [22]. This increases the number of parameters to estimate, but is more powerful. If variance compensation is also desired, the same transformation matrices as used for the means is applied, so as to keep the number of parameters manageable [23].

One of the major drawbacks of applying MLLR adaptation to noise-corrupted environments is that the transformation matrices are estimated given an alignment between frames and states, the initial estimate of which may be poor. This will yield poor transformation matrices and, thus, poor performance. Various solutions have been proposed to alleviate this problem [24].

Another limitation of the MLLR framework arises due to the intrinsic non-linearities of the noise compensation process. To model this transformation well, a sufficient number of regression classes must be used. Unfortunately, there is not always sufficient adaptation material to allow this [23].

3.4 Non-Linear Transformation

This prompted investigations into the possibility of using non-linear mappings to relate speech parameters (features or models) in training and operating environments. In the feature space, the metamorphic algorithm provides a way to estimate a piecewise linear transformation between the associated spaces [25]: for each speech unit, the optimal linear transformation is estimated in the least-squares sense. Linear input networks and Gamma networks have also been used in conjonction with gradient descent on the connectionist parameters [26]. In addition, some of these techniques, such as parallel hidden networks, have been applied in the model space as well [26].

Finally, we include in this category a number of approaches based on linear approximations to non-linear transformations. In [27], for example, a vector Taylor series approximation is used to relate noise vectors, noisy speech vectors, and the associated statistical distributions. A similar approach based on the Jacobian matrix has recently been described in [28].

4 PERSPECTIVES

The techniques mentioned above represent a large and diverse collection, ranging from generally applicable approaches to extremely narrow solutions, with a host of trade-off points to consider in the selection process. Complicating the issue is the fact that sometimes the methods have been evaluated under overly restrictive conditions, such as additive white Gaussian noise, which makes it difficult to predict their performance in more realistic environments.

Not surprisingly, there has been much emphasis lately on combining some of these techniques, in an effort to harness the benefits of each. For example, several ideas previously developed in the context of parallel model combination, bias removal, and stochastic matching are integrated in [29], and [30] combines stochastic matching, maximum a posteriori re-estimation, and transfer vector interpolation.

Also of interest is the fact that several independent investigations (see, e.g., [31], [32]) have integrated prior information into linear regression. This was shown to achieve good performance for a wide range of adaptation set sizes, effectively making fast adaptation practical [31].

One outstanding issue that should probably be addressed more aggressively is the fact that noise corruption leads to non-Gaussian corrupted speech distributions, even in the case of Gaussian speech and noise sources [21]. This raises disturbing questions as to the validity of many theoretical developments.

It also implies that, barring the use of non-Gaussian distributions (which ones?), the accurate representation of the corrupted distributions requires an adjustment in the model itself, not just its parameters. For example, additional components may be necessary in an underlying Gaussian mixture model, which not only increases

the computational load but also requires suitable methods to add these components. As an example, an algorithm was presented in [33] to control the model complexity through an information criterion, the minimum description length (MDL) principle. It is likely that further insights into the non-Gaussian modeling problem will steadily emerge over the coming years.

REFERENCES

[1] B.H. Juang, "Speech Recognition in Adverse Environments," *Comput. Speech Language*, Vol. 5, pp. 275–294, 1991.

[2] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Boston, MA: Kluwer Acad. Publ., 1993.

[3] Y. Gong, "Speech Recognition in Noisy Environments: A Survey," *Speech Comm.*, Vol. 16, pp. 261–291, 1995.

[4] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, Boston, MA: Kluwer Acad. Publ., 1996.

[5] L. Neumeyer and M. Weintraub, "Probabilistic Optimum Filtering for Robust Speech Recognition," in *Proc. ICASSP*, Vol. I, pp. 417–420, 1994.

[6] Y.M. Cheng, *et al.*, "Statistical Signal Mapping: A General Tool for Speech Signal Processing," in *Proc. 6th IEEE Worksh. Stat. Sig., Array Proc.*, pp. 436–439, 1992.

[7] Y. Anglade, *et al.*, "Speech Discrimination in Adverse Conditions Using Acoustic Knowledge and Selectively Trained Neural Networks," in *Proc. ICASSP*, Vol. II, pp. 279–282, 1993.

[8] Y. Gao and J.-P. Haton, "A Hierarchical LPNN Network for Noise Reduction and Noise Degraded Speech Recognition," in *Proc. ICASSP*, Vol. II, pp. 89–92, 1994.

[9] R.C. Rose, *et al.*, "Integrated Models of Signal and Background with Application To Speaker Identification in Noise," *IEEE Trans. SAP*, Vol. 2, No. 2, pp. 245–257, 1994.

[10] A. Nadas, *et al.*, "Speech Recognition Using Noise–Adaptive Prototypes," *IEEE Trans. ASSP*, Vol. 37, No. 10, pp. 1495–1502, 1989.

[11] M.J.F. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. Thesis, Cambridge U. September 1995.

[12] P.C. Woodland, *et al.*, "Improving Environmental Robustness in Large Vocabulary Speech Recognition," in *Proc. ICASSP*, Vol. I, pp. 65–68, 1996.

[13] M.G. Rahim and B.H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," *IEEE Trans. SAP*, Vol. 4, No. 1, pp. 19–30, 1996.

[14] F. Liu, *et al.*, "Signal Processing for Robust Speech Recognition," in *Proc. ARPA Worksh. Human Lang. Tech.*, pp. 309–314, 1994.

[15] Y. Zhao, "An Acoustic–Phonetic–Based Speaker Adaptation Technique Improving Speaker–Independent Continuous Speech Recognition," *IEEE Trans. SAP*, Vol. 2, No. 3, pp. 380–394, 1994.

[16] M. Tonomura, *et al.*, "Speaker Adaptation Based on Transfer Vector Field Smoothing Using MAP Estimation," in *Proc. ICASSP*, pp. 688–691, 1995.

[17] J.L. Gauvain and C.H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. SAP*, Vol. 2, No. 2, pp. 291–298, 1994.

[18] A.C. Surendran, *Maximum–Likelihood Stochastic Matching Approach to Non-Linear Equalization for Robust Speech Recognition*, Ph.D. Thesis, Rutgers U., May 1996.

[19] A. Sankar and C.H. Lee, "A Maximum–Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Trans. SAP*, Vol. 4, No. 2, pp. 190–202, 1996.

[20] P.J. Moreno, *et al.*, "Multivariate–Gaussian–Based Cepstral Normalization for Robust Speech Recognition," in *Proc. ICASSP*, pp. 137–140, 1995.

[21] J.P. Openshaw and J.S. Mason, "On the Limitations of Cepstral Features in Noise," in *Proc. ICASSP*, Vol. II, pp. 49–52, 1994.

[22] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Comput. Speech Language*, Vol. 9, No. 2, pp. 171–186, 1995.

[23] M.J.F. Gales and P.C. Woodland, "Mean and Variance Adaptation Within the MLLR Framework," *Comput. Speech Language*, Vol. 10, pp. 249–264, 1996.

[24] T. Anastasakos, *et al.*, "A Compact Model for Speaker–Adaptive Training," in *Proc. 4th ICSLP*, pp. 1137–1140, 1996.

[25] J.R. Bellegarda, *et al.*, "The Metamorphic Algorithm: A Speaker Mapping Approach to Data Augmentation," *IEEE Trans. SAP*, Vol. 2, No. 3, pp. 413–420, 1994.

[26] J. Neto, *et al.*, "Speaker–Adaptation for Hybrid HMM–ANN Continuous Speech Recognition System," in *Proc. EuroSpeech'95*, pp. 2171–2174, 1995.

[27] P.J. Moreno, *et al.*, "A Vector Taylor Series Approach for Environment–Independent Speech Recognition," in *Proc. ICASSP*, pp. 835–838, 1997.

[28] S. Sagayama, *et al.*, "Jacobian Approach To Fast Acoustic Model Adaptation," in *Proc. ICASSP*, pp. 835–838, 1997.

[29] M. Afify, *et al.*, "A Unified Maximum Likelihood Approach to Acoustic Mismatch Compensation: Application to Noisy Lombard Speech Recognition," in *Proc. ICASSP*, pp. 839–842, 1997.

[30] J.T. Chien, *et al.*, "Improved Bayesian Learning of Hidden Markov Models for Speaker Adaptation," in *Proc. ICASSP*, pp. 1027–1030, 1997.

[31] V. Digalakis and L. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods," *IEEE Trans. SAP*, Vol. 4, No. 3, pp. 294–300, 1996.

[32] G. Zavaliagkos, *et al.*, "Maximum A Posteriori for Large Scale HMM Recognizers," in *Proc. ICASSP*, Vol. I, pp. 725–728, 1996.

[33] K. Shinoda and T. Watanabe, "Speaker Adaptation with Autonomous Model Complexity Control by MDL Principle," in *Proc. ICASSP*, pp. 717–720, 1996.