ROLES OF HIGH-FIDELITY ACOUSTIC MODELING IN ROBUST SPEECH RECOGNITION

Li Deng

Microsoft Research, One Microsoft Way, Redmond, WA 98052
deng@microsoft.com

ABSTRACT

In this paper I argue that high-fidelity acoustic models have important roles to play in robust speech recognition in face of a multitude of variability ailing many current systems. The discussion of high-fidelity acoustic modeling is posited in the context of general statistical pattern recognition, in which the probabilistic-modeling component that embeds partial, imperfect knowledge is the fundamental building block enabling all other components including recognition error measure, decision rule, and training criterion. Within the session's theme of acoustic modeling and robust speech recognition, I advance my argument using two concrete examples. First, an acoustic-modeling framework which embeds the knowledge of articulatory-like constraints is shown to be better able to account for the speech variability arising from varying speaking behavior (e.g., speaking rate and style) than without the use of the constraints. This higher-fidelity acoustic model is implemented in a multi-layer dynamic Bayesian network and computer simulation results are presented. Second, the variability in the acoustically distorted speech under adverse environments can be more precisely represented and more effectively handled using the information about phase asynchrony between the un-distorted speech and the mixing noise than without using such information. This high-fidelity, phase-sensitive acoustic distortion model is integrated into the same multi-layer Bayesian network but at separate, causally related layers from those representing the speaking-behavior variability. Related experimental results in the literature are reviewed, providing empirical support to the significant roles that the phase-sensitive model plays in environment-robust speech recognition.

Index Terms— acoustic modeling, noise robustness, speaking behavior, variability, high fidelity, generative modeling, phase asynchrony, dynamic Bayesian network

1. INTRODUCTION

Statistical pattern recognition, including modern methods in automatic speech recognition (ASR), has been experiencing a long history of development. In virtually all useful applications of statistical pattern recognition (e.g., ASR, handwriting recognition, machine translation, image recognition, etc.), the tasks are complex and difficult, for which no straightforward physical principles can easily provide satisfactory solutions. As a result, researchers seek to use whatever reliable and relevant sources of knowledge, albeit their imperfection, to make the statistical decisions that are "optimal" given such partial and possibly vague knowledge. The decisions often come down to the minimization of specific types of decision error measures; e.g., the number of errors in a string of words (sentence) or word error rate in ASR.

Given the current state of statistical pattern recognition, ASR in particular, as outlined above, the central issues to be addressed by the research community can be summarized as follows. First, how should we construct statistical models that embed partial, imperfect knowledge in a probabilistic fashion? Confined within this session's theme of acoustic modeling and robust ASR, this issue becomes: How to use our incomplete understanding of the human speech process, of the variability of speech acoustics, and of the nature of the speech distortion under adverse acoustic environments to build and refine probabilistic models for the observed, highly variable speech measurements? Second, how should we specify the performance of statistical pattern recognition? And for ASR, how do we define and measure ASR errors? Should we discount errors for semantically irrelevant functional words, and should we look for phone string errors? Third, how should we define the training's objective functions that guide the learning of the free parameters in the statistical models? In the case of hidden Markov models (HMMs) for speech acoustics, what is the most appropriate objective function (e.g., minimum classification error, maximum mutual information, minimum phone/word error, etc.) for HMM learning? And given the training objective function, how to efficiently optimize them? The key problems here are 1) consistency between the decision error measure and the training objective function, and 2) generalization capability of the training objective function to the unseen test data drawn from each the same or different statistical distributions. And finally, how to design the decision rule for the recognition/decoding process, and how to efficiently implement the decision rule? The key issues here is, again, consistency between the performance

The author thanks Hermann Ney, Alex Acero, Dong Yu, Jinyu Li, Jasha Droppo for insightful discussions and selected experimental results summarized in this paper.

measure and the decision rule.

Among all the above central ingredients in the modern statistical methods in ASR (possibly in other statistical pattern recognition problems also), the first ingredient, that of building high-quality probabilistic models for speech, is arguably the most difficult and important one. Substantial amounts of my personal research effort in the past have been devoted to this area. In this paper I will use two concrete examples to illustrate the key roles of high-fidelity acoustic modeling in robust ASR, where robustness refers to that against all types and sources of variability. For the remaining three key elements which will be the focus of this paper, the readers are referred to a number of relatively recent literature [26, 5, 14, 23, 33, 12, 18, 32].

2. ISSUES IN HIGH-FIDELITY ACOUSTIC MODELING

There have been many types of statistical models for speech acoustics developed over the past three decades. They can be broadly classified into two main categories: 1) generative models, and 2) discriminative models. Generative speech recognizers (e.g., [16, 29, 6]), such as those based on mixture models, HMMs, and stochastic segment models, rely on a learned model of the joint probability distribution of the observed acoustic features and the corresponding speech class membership. They use this joint-probability characterization to perform the decision making task based on the posterior probability of the class computed by Bayes rule. In contrast, discriminative speech recognizers (e.g., [30]), such as those based on maximum entropy models, neural networks, and conditional random fields, directly employ the speech class posterior probability or the related discriminant function. The discriminative recognizer design philosophy is the basis of a wide range of popular machine learning methods, where some known deficiencies of the HMM are addressed by applying direct discriminative learning and hence replacing the need for a probabilistic generative model by a set of flexibly selected, overlapping features. Since the conditioning is made on the feature sequence and these features can be designed with long-contextual-span properties, the conditional-independence assumption made in the HMM is conceptually alleviated – provided that proper features can be constructed. How to design such features is a challenging research direction and it becomes a critical factor for the potential success of the structured discriminative approach. On the other hand, local features can be more easily designed that are appropriate for the generative approach and many effective local features have been established for speech recognition. Despite the complexity of estimating joint distributions when the sole purpose is discrimination, the generative approach has important advantages of facilitating knowledge incorporation and of conceptually straightforward analyses of the recognizer's components and their interactions.

Current state of acoustic modeling in ASR is that the capabilities and limitations associated with both generative and discriminative approaches discussed above are compromised, leading to practical recognition frameworks where simplistic joint-distribution models (such as HMMs) are established to characterize the statistical properties of speech, with the complexity lower than what is required to accurately generate samples from the true distribution. In order to make such low-complexity, low-fidelity generative models discriminate well, it requires parameter learning methods that are discriminative in nature to overcome the limitation in the simplistic HMM structure. This is in contrast to the generative approach of fitting the intra-class data as conventional maximum likelihood based methods intend to accomplish. This type of practical frameworks has been applied to much of the recent work in speech recognition research, where HMMs are used as the low-complexity joint distribution for the local acoustic feature sequences of speech and the corresponding underlying linguistic sequences of sentences, words, or phones.

For advancing the state of the art in acoustic modeling for robust ASR, it is this author's belief that both the generative and discriminative modeling approaches require acoustic models with higher fidelity than the common approaches seen today, and that their respective strengths as discussed above may be combined to achieve greater effectiveness. Current state of HMM-based acoustic modeling has intended and is able to capture only a subset of the tremendous variability in speech acoustics, often in an isolated, non-systematic way. To achieve true robustness in ASR, we need to handle all sources of the variability, including 1) pronunciation variability; 2) variability due to accent and dialect; 3) variability due to prosodic and phonetic context; 4) variability due to speaking behavior (e.g., style and rate); 5) variability due to the adverse speaking condition that affects articulation; 6) variability due to noisy acoustic environment; 7) transducer variability and distortions; and 8) transmission channel variability and distortions.

How to systematically handle all these types of speech variability in the discriminative modeling framework appears to be less straightforward than in the generative modeling framework, partly because the much longer history of development of the latter. Even within the generative modeling framework, the HMM framework in particular, much research remains to represent and to compensate for all the main sources of variability in a principled and systematic way. After presenting in the next section a general, probabilistic framework to characterize the multi-layered, causal mechanisms as a type of high-fidelity speech model, I will use two concrete sub-problems within this framework in the following two sections to illustrate how to account for two specific types of variability that is difficult for the conventional techniques.

3. A MULTI-LAYER DYNAMIC BAYESIAN NETWORK MODEL FOR SPEECH ACOUSTICS

The modern machine learning tool, dynamic Bayesian network [4], is a powerful probabilistic framework to represent underlying data-generation mechanisms and the associated variability for sequential data such as acoustic feature sequences of speech. One particular form of the dynamic Bayesian network that is capable of accounting for many (not all) types of the variability discussed in the preceding section is illustrated in Fig. 1. The S-layer (top) in Fig. 1 represents the temporal dynamics of discrete linguistic units. The t-layer represents the dynamics in the segmental phonetic targets that are continuous valued and whose statistical distributions are correlated with the discrete-valued linguistic units. The z-layer represents the "articulation-like" temporal dynamics that are driven by the segmental targets, where the continuity constraint in the z values across segment boundaries and the limits on how fast the z values can change over time jointly accounts for the variability due to phonetic context and that due to speaking behavior. Examples of "articulation-like" z vectors include reduced-dimension articulatory parameters (after principal component analysis), vocal tract area functions, and vocal tract resonances (formants and the associated bandwidths). The o-layer represents the non-distorted speech dynamics that are generated causally from the articulation-like dynamics. And finally, the y-layer represents the observed speech dynamics after environmental distortion. The adverse environmental condition is characterized, in the general term, by two sets of parameters: the time-varying parameters for additive noise are represented by n-layer, which is controlled by discrete variables (N-layer) from different noise types, and the time-invariant parameter convolutive channel distortion is represented by the fixed h variable.

The multi-layer dynamic Bayesian network model shown in Fig. 1, while intended to be comprehensive, has not been able to account for the variability due to the adverse speaking condition that affects articulation (e.g., stress speech, Lombard effect, or hyper-articulated speech). In order to incorporate this effect, a feedback from the n-layer or y-layer to z-layer can be added. However, this addition would make inference and estimation problems in the expanded dynamic Bayesian network model significantly more complex, which will not be addressed here. Further, other types of variability (pronunciation, accent and dialect, and prosodic variations) have not been carefully represented in the Bayesian network model of Fig. 1. One possible way to handle these types of variability is to expand the S-layer into multiple-tiers, improving the single-tier "beads-on-the-string" "pronunciation model" in Fig. 1 to a non-linear, multi-tier model as proposed and initially implemented in [28, 7, 35], with the possible mathematical representation like "factorial HMM" developed in [27].

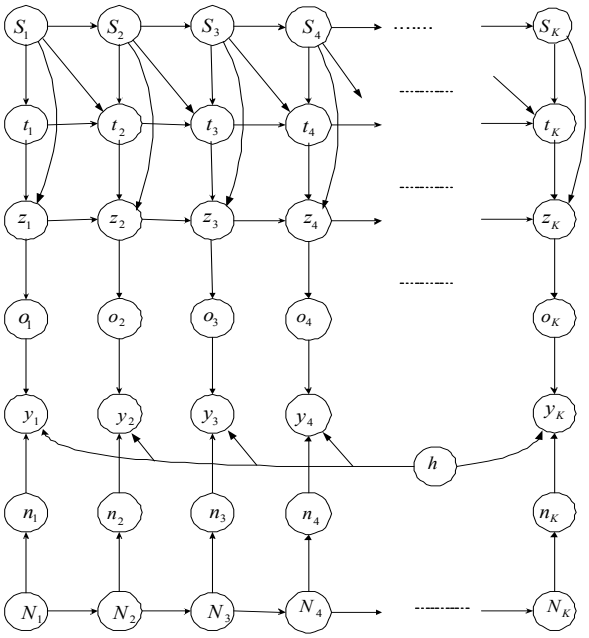

Fig. 1. A multi-layer dynamic Bayesian network model from the linguistic units (S-layer) to distorted speech acoustics (y-layer). Intermediate layers include the phonetic target model (t-layer), articulation-like dynamic model (z-layer), (clean) acoustic observation model (o-layer), and the distorted acoustic observation model (y-layer). Time-varying parameters (noise vectors) in the environment-distortion model are represented by n-layer, which is controlled by discrete variables (N-layer) representing different noise types. The time-invariant parameter (channel vector) is represented by the h variable.

4. ACCOUNTING FOR VARIABILITY DUE TO SPEAKING BEHAVIOR

The main underlying cause for the variability in the acoustic observation of speech due to factors related to speaking behavior lies in the "hidden" domain of un-observed articulation and its control. In the multi-layer dynamic Bayesian network model shown in Fig. 1, this cause is naturally represented in the combined hidden S-layer, t-layer, and z-layer, where the latter represents the hidden "articulation-like" dynamics driven by the segmental, S-dependent, targets (t-layer) which functionally serve as the input or the "control" signal to the "articulatory" system. Since the z vector is intended to represent the physical, "articulatory" structure with inertia properties, the continuity constraint is naturally imposed that limits the movement pattern of the z vectors over time both within and across the discrete segment boundaries. The resulting constrained movement pattern in the z vectors as modeled by the conditional dependencies represented in the t-layer and z-layer of the dynamic Bayesian network model shown in Fig.

1 accounts for the speech variability due to both the speaking rate and style aspects of (passive) speaking behavior. The varying speaking style can be represented by statistical distributions in the parameter of "time constant" that governs the dynamic behavior in the z-layer.

Several different implementations of the combined hidden S-layer, t-layer, and z-layer dynamics can be found in [6], with detailed mathematical descriptions on the conditional dependencies among these layers as well as on the conditional dependency to the o-layer for the un-distorted observed speech acoustics. Positive phonetic recognition results have been reported in [8, 9]. In the reminder of this section, I will show some representative computer simulation results that demonstrate major dynamic properties (e.g., target undershooting or reduction) in one specific implementation where vocal tract resonances (VTR) or formants are used as the continuous-valued z vector. These results are further compared with the corresponding results on the direct measurements of the corresponding properties in the acoustic-phonetic literature. (More detailed information about these simulation results and the related model can be found in [8, 6].)

To illustrate VTR frequency or formant target undershooting, we first show the spectrogram of three renditions of a three-segment /iy aa iy/ in Fig. 2. From left to right, the speaking rate increases and speaking effort decreases, with the durations of the /aa/'s decreasing from approximately 230 msec to 130 msec. Formant target undershooting for f_1 and f_2 is clearly visible in the spectrogram, where automatically tracked formants are superimposed (as the solid lines in Fig. 2 to aid identification of the formant trajectories.

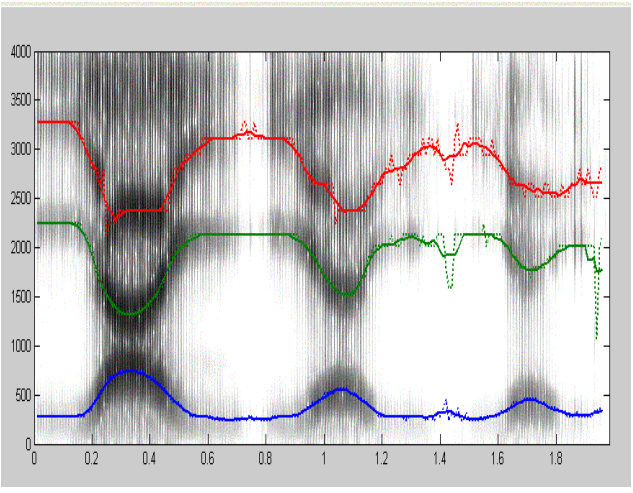

Fig. 2. Spectrogram of three renditions of /iy aa iy/ by the author, with an increasingly higher speaking rate and increasingly lower speaking efforts (i.e., the speaking style becomes more casual). The horizontal label is time (in seconds), and the vertical one is frequency (in Hz).

4.1. Effects of "time constant" parameter on reduction in model simulation

The same kind of target undershooting for f_1 and f_2 as in Fig. 2 is exhibited in the combined t-layer and z-layer model prediction, shown in Fig. 3, where we also illustrate the effects of the "time constant" parameter, γ, on the magnitude of formant undershooting or reduction. The model prediction is for the f_1 and f_2 of the z vector. Figs. 3a, b, and c correspond to the use of the "time constant" parameter value (the same for each formant vector component) set at $\gamma = 0.85, 0.75$ and 0.65, respectively, where in each plot the slower /iy aa iy/ sounds (with the duration of /aa/ set at 230 msec or 23 frames) are followed by the faster /iy aa iy/ sounds (with the duration of /aa/ set at 130 msec or 13 frames). f_1 and f_2 targets for /iy/ and /aa/ are set appropriately in the model also. Comparing the three plots, we obtain the model's quantitative prediction: The magnitude of reduction decreases as the γ value decreases.

In Figs. 4a, b, and c, we show the same model prediction as in Fig. 3 but for different sounds /iy eh iy/, where the targets for /eh/ are much closer to those of the adjacent sound /iy/ than in the previous case for /aa/. As such, the absolute amount of reduction becomes smaller. However, the same effect of the "time constant" parameter's value on the magnitude of reduction is shown as for the previous sounds /iy aa iy/, except the effect becomes less pronounced.

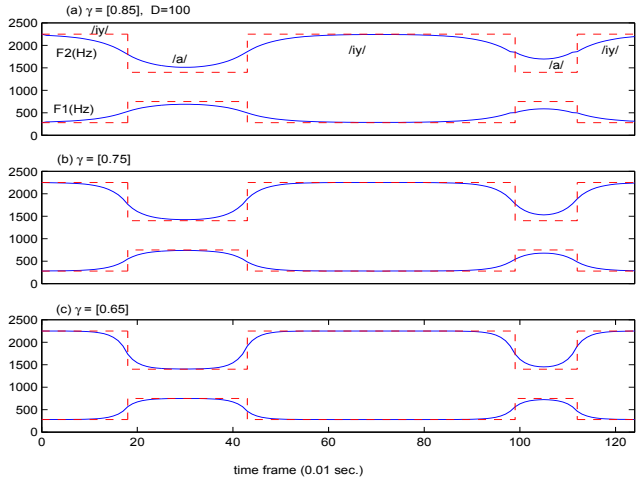

Fig. 3. f_1 and f_2 formant or VTR frequency trajectories produced from the model for a slow /iy aa iy/ followed by a fast /iy aa iy/. (a), (b), and (c) correspond to the use of the stiffness parameter values of $\gamma = 0.85, 0.75$ and 0.65, respectively. The amount of formant undershooting or reduction during the fast /aa/ is decreasing as the γ value decreases. The dashed lines indicate the formant target values and their switch at the segment boundaries.

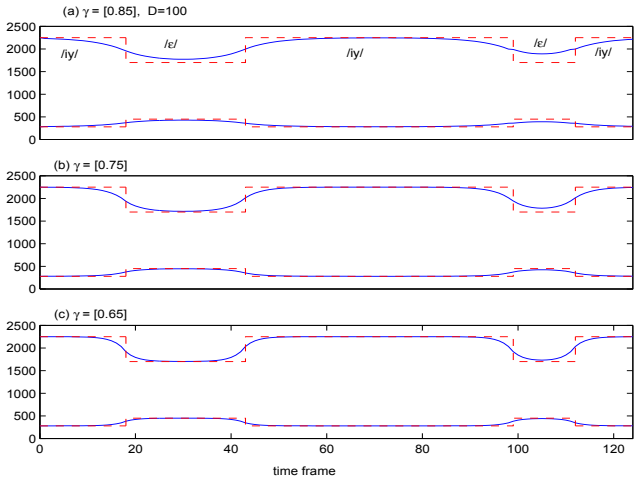

Fig. 4. Same as Fig. 3 except for the /iy eh iy/ sounds. Note that the f_1 and f_2 target values for /eh/ are closer to /iy/ than those for /aa/.

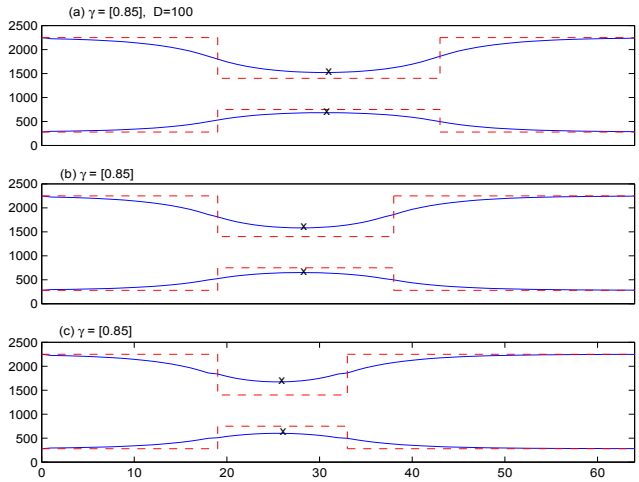

Fig. 5. f_1 and f_2 formant trajectories produced from the model for three different durations of /aa/ in the /iy aa iy/ sounds: (a) 25 frames (250 ms), (b) 20 frames, and (c) 15 frames. The same γ value of 0.85 is used. The amount of target undershooting increases as the duration is shortened or the speaking rate is increased. Symbols 'x' mark the f_1 and f_2 formant values at the central portions of vowels of /aa/.

4.2. Effects of speaking rate on reduction in model simulation

In Fig. 5, we show the effects of speaking rate, measured as the inverse of the sound segment's duration, on the magnitude of reduction of undershooting. Sub-plots (a), (b), and (c) correspond to three decreasing durations of the sound /aa/ in the /iy aa iy/ sound sequence. They illustrate an increasing amount of the reduction with the decreasing duration or increasing speaking rate. Symbol 'x' in Fig. 5 indicates the f_1 and f_2 formant values at the central portions of vowels /aa/, which are predicted from the model and are used to quantify the magnitude of reduction. These values (separately for f_1 and f_2) for /aa/ are plotted against the inversed duration in Fig. 6, together with the corresponding values for /eh/ in the /iy eh iy/ sound sequence. The most interesting observation is that as the speaking rate increases, the distinction between vowels /aa/ and /eh/ gradually diminishes if their static formant values extracted from the dynamic patterns are used as the sole measure for the difference between the sounds. We refer to this phenomenon as "static" sound confusion induced by increased speaking rate (or/and by a greater degree of sloppiness in speaking).

4.3. Comparisons with formant measurement data

The "static" sound confusion between /aa/ and /eh/ quantitatively predicted by the model as shown in Fig. 6 is consistent with the formant measurement data published in [31], where thousands of natural sound tokens were used to investigate the relationship between the degree of formant undershooting and speaking rate. We re-organized and re-plotted the raw data from [31] in Fig. 7, in the same format as Fig. 6. While the measures of speaking rate differ between the measurement data and model prediction and cannot be easily converted to each other, the overall results are generally consistent with each other. The similar trend for the greater degree of "static" sound confusion as speaking rate increases is shown clearly from both the measurement data (Fig. 7) and the model prediction (Fig. 6).

4.4. Model prediction of vocal tract resonance trajectories for real speech utterances

We have used the expected VTR trajectories computed from the model to predict actual VTR frequency trajectories for real speech utterances from the TIMIT database. Only the phone identities and their boundaries are input to the model for the prediction, and no use is made of speech acoustics. Given the phone sequence in any utterance, we first break up the compound phones (affricates and diphthongs) into their constituents. Then we obtain the initial VTR target values based on limited context dependency by table lookup. Then automatic and iterative target adaptation is performed for each phone-like unit based on the difference between the results of a VTR tracker and the VTR prediction from the model output. These target values are provided not only to vowels, but also to consonants for which the resonance frequency targets are used with weak or no acoustic manifestation. The converged target values, together with the phone boundaries provided from the TIMIT database, form the input to the z-layer of the model and the output of the z-layer gives the predicted VTR frequency trajectories.

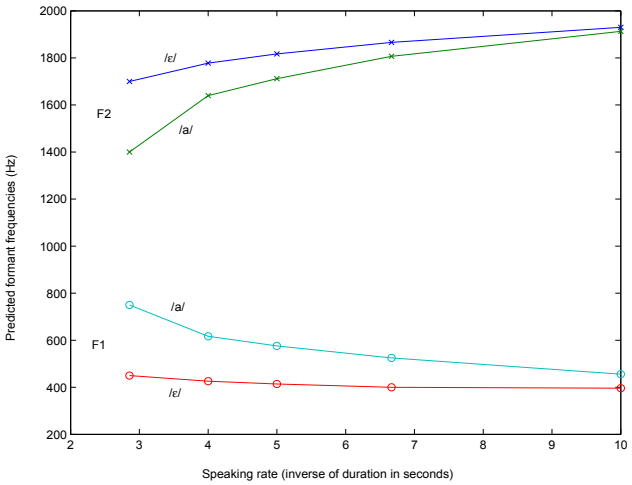

Fig. 6. Relationship, based on model prediction, between the f_1 and f_2 formant values at the central portions of vowels and the speaking rate. Vowel /aa/ is in the carry-phrase /iy aa iy/, and vowel /eh/ in /iy eh iy/. Note that as the speaking rate increases, the distinction between vowels /aa/ and /eh/ measured by the difference between their static formant values gradually diminishes. The same γ value of 0.9 is used in generating all points in the figure.

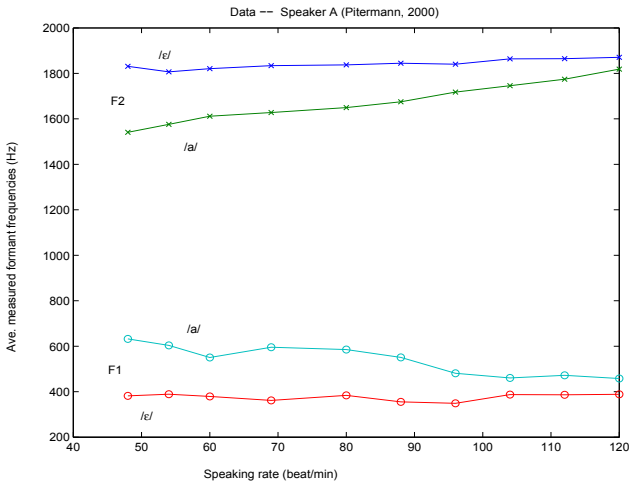

Fig. 7. The formant measurement data from literature are reorganized and plotted, showing similar trends to the model prediction under similar conditions.

Three example utterances from TIMIT (SI1039, SI1669, and SI2299) are shown in Figs. 8-10. The step-wise dashed lines ($f_1/f_2/f_3/f_4$) are the target sequences as inputs, and the continuous lines ($f_1/f_2/f_3/f_4$) are the outputs of the z-layer model as the predicted VTR frequency trajectories. To facilitate assessment of the accuracy in the prediction, the inputs and outputs are superimposed on the spectrograms of these utterances, where the true resonances are shown as the dark bands. For the majority of frames, the output either coincides or is close to the true VTR frequencies, even though no acoustic information is used. Also, comparing the input and output, we observe a relatively mild degree of target undershooting or reduction in these and many other TIMIT utterances that we have examined.

4.5. Section summary

The computer simulation results presented in this section show that the combined model for the t-layer and z-layer which embeds the knowledge of articulatory-like constraints can effectively account for the speech variability due to a range of speaking behavior. The conventional HMMs, which may not naturally use such constraints, have difficulties in capturing this type of speaking-behavior variability in a parsimonious manner. So far, the most comprehensive implementation and evaluation of the model (where VTRs are used as the z vector) have been applied to the standard phonetic recognition task of TIMIT, a relatively small task due mainly to the high computational cost in decoding (not in training). The results presented in [9] show the significantly higher phonetic recognition rate (75.1%) than a state-of-the-art HMM system (71.4%). Error analysis shows that the improvements are most significant in the sonorant class, followed by the stop-consonant class. No improvement is observed in the fricative-consonant class. This is in accord with our expectation since the model component design has a greater degree of precision (i.e., higher "fidelity") for the VTR dynamics and its mapping to the cepstral features as the acoustic observation for the sonorant class of speech sounds. The performance improvement for the stop-consonant class is likely due to the better modeling of vocalic portions of VTR transitions from stop to vowel and from vowel to stop.

5. ACCOUNTING FOR VARIABILITY DUE TO ADVERSE ACOUSTIC ENVIRONMENT

5.1. Introduction

In this section, we focus on another major type of speech variability, that due to the adverse acoustic environment with both additive and convolutive (with short-term impulse responses) distortions. (The distortion caused by convolutive distortion with long-term impulse responses or reverberation will not be discussed here.) Handling this type of variability has high practical value since it is directly related to the deployment of speech recognizers. Environment robustness in speech recognition remains an outstanding and difficult problem despite many years of research and investment (e.g., [1, 17, 24, 19, 2, 36, 15, 13, 25, 22, 20]). The difficulty arises due to many possible types of distortions, including a varying degree of additive and convolutive distortions and

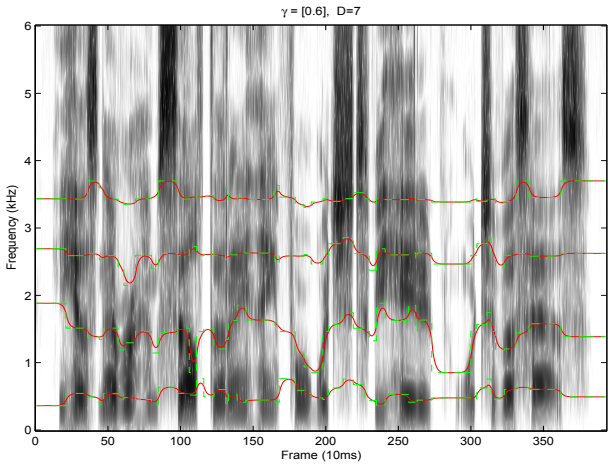

Fig. 8. The $f_1/f_2/f_3/f_4$ VTR frequency trajectories (smooth lines) generated from the model for VTR target filtering using the phone sequence and duration of a speech utterance (SI1039) taken from the TIMIT database. The target sequence is shown as stepwise lines, switching at the phone boundaries labeled in the database. They are superimposed on the utterance's spectrogram. The utterance is "*He has never, himself, done anything for which to be hated – which of us has* ".

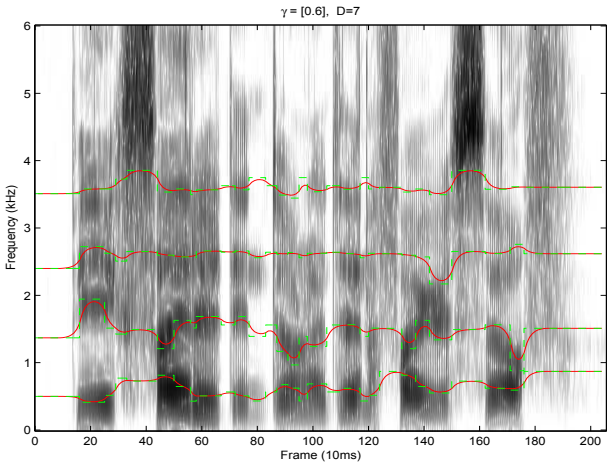

Fig. 9. Same as Fig. 8 except with another utterance "*Be excited and don't identify yourself*" (SI1669).

their mixes, which are not easy to predict accurately during recognizer deployment. As a result, the speech recognizer trained using clean speech often degrades its performance significantly when used under noisy environments if no environment-robustness strategy is applied.

The portions of the high-fidelity acoustic model handling environment robustness in the multi-layered Bayesian network of Fig. 1 are in the combined o-layer, N-layer, n-layer, h-variable, and y-layer, where the y-layer represents observational feature sequences of distorted speech and all other layers (including the clean-speech o-layer) are hidden. The mathematical representation for the conditional dependency of the y-layer on the o-layer, n-layer, and the h-value in the Bayesian network of Fig. 1 defines the "acoustic model for environmental distortion". This is a parsimonious, parametric model, since the model is characterized by only the noise and channel parameters. (This is in contrast to the conventional MLLR-type, data-driven distortion model where many transformation matrices are used.)

Traditionally, the acoustic model for environmental distortion ignores the phase asynchrony between the clean speech and the mixing noise [1, 24]. Such a "low-fidelity" model has been improved, over the past several years, to achieve "higher fidelity" that removes the earlier simplifying assumption by including random phase asynchrony in the distortion model [2, 36, 10, 11, 34, 21]. Since this gives a most fitting example to illustrate the roles of high-fidelity modeling in robust ASR, in this section, we first give detailed derivation of the phase-sensitive model (with both the deterministic and probabilistic versions). Then, we summarize the existing experimental results that illustrate performance gain by moving from "low-fidelity", phase-insensitive model to the "high-fidelity" phase-sensitive model, and offer insight to understanding the roles of incorporating the phase information by analyzing the experimental results.

5.2. The phase-sensitive model of environmental distortion — Deterministic version

In this subsection, we derive the phase-sensitive model in the log filter-bank domain. (This can be easily extended to the cepstral domain, which will not be included in this paper.) Using the discrete-time, linear system model for the acoustic distortion in the time domain, we have the well-known relationship among the noisy speech ($y(t)$), clean speech ($x(t)$), additive noise ($n(t)$), and the impulse response of the linear distortion channel ($h(t)$):

$$y(t) = x(t) * h(t) + n(t).$$

In the frequency domain, the equivalent relationship is

$$Y[k] = X[k]H[k] + N[k], \tag{1}$$

where k is the frequency-bin index in DFT given a fixed-length time window, and $H(k)$ is the (frequency-domain) transfer function of the linear channel.

The power spectrum of the noisy speech can then be obtained from the DFT in Eq. 1 by

$$\begin{aligned} |Y[k]|^2 &= |X[k]H[k] + N[k]|^2 \\ &= |X[k]|^2|H[k]|^2 + |N[k]|^2 + (X[k]H[k])(N[k])^* \\ &\quad +(X[k]H[k])^* N[k] \\ &= |X[k]|^2|H[k]|^2 + |N[k]|^2 + 2|X[k]||H[k]||N[k]|\cos\theta_k, \end{aligned} \tag{2}$$

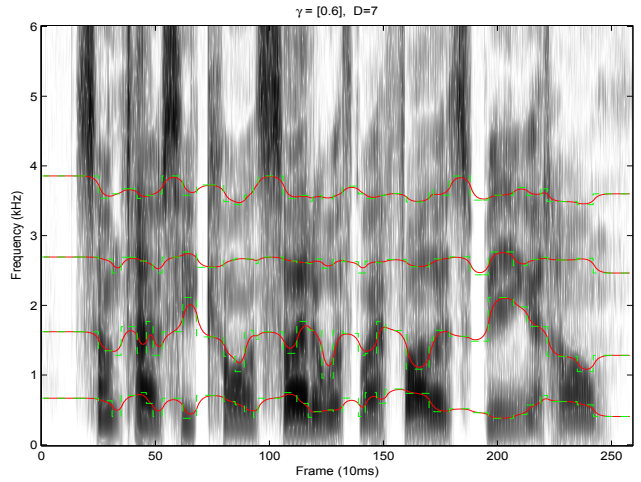

Fig. 10. Same as Fig. 8 except with the third utterance "*Sometimes, he coincided with my father's being at home*" (SI2299).

where θ_k denotes the (random) phase angle between the two complex variables $N[k]$ and $(X[k]H[k])$. Eq. 2 incorporates the phase relationship between the (linearly filtered) clean speech and the additive corrupting noise in the speech distortion process. It is noted that in the traditional, phase-insensitive models for acoustic distortion, the last term in Eq. 2 has been assumed to be zero. This is correct only in expected sense. The phase-sensitive model presented here based on Eq. 2 with non-zero instantaneous values in the last term removes this commonly made but un-realistic assumption.

After applying a set of Mel-scale filters (L in total) to the spectrum $|Y[k]|^2$ in the frequency domain, where the l^{th} filter is characterized by the transfer function $W_k^{(l)} \geq 0$ (where $\sum_k W_k^{(l)} = 1$), we obtain a total of L Mel-filter-bank energies of

$$\begin{aligned}\sum_k W_k^{(l)}|Y[k]|^2 &= \sum_k W_k^{(l)}|X[k]|^2|H[k]|^2 + \sum_k W_k^{(l)}|N[k]|^2 \\ &+ 2\sum_k W_k^{(l)}|X[k]||H[k]||N[k]|\cos\theta_k, \end{aligned} \quad (3)$$

with $l = 1, 2, ..., L$.

Denoting the various filter-bank energies in Eq. 3 by

$$\begin{aligned}|\tilde{Y}^{(l)}|^2 &= \sum_k W_k^{(l)}|Y[k]|^2, \\ |\tilde{X}^{(l)}|^2 &= \sum_k W_k^{(l)}|X[k]|^2, \\ |\tilde{N}^{(l)}|^2 &= \sum_k W_k^{(l)}|N[k]|^2, \end{aligned} \quad (4)$$

and

$$|\tilde{H}^{(l)}|^2 = \frac{\sum_k W_k^{(l)}|X[k]|^2|H[k]|^2}{|\tilde{X}^{(l)}|^2},$$

we simplify Eq. 3 to

$$|\tilde{Y}^{(l)}|^2 = |\tilde{X}^{(l)}|^2|\tilde{H}^{(l)}|^2 + |\tilde{N}^{(l)}|^2 + 2\alpha^{(l)}|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|, \quad (5)$$

where we define the "phase factor" as

$$\alpha^{(l)} \equiv \frac{\sum_k W_k^{(l)}|X[k]||H[k]||N[k]|\cos\theta_k}{|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|}. \quad (6)$$

Since $\cos\theta_k \leq 1$, we have

$$|\alpha^{(l)}| \leq \frac{\sum_k W_k^{(l)}|X[k]||H[k]||N[k]|}{|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|}.$$

The right-hand side is the normalized inner product of vectors $\bar{N}$ and $\bar{X}^H$, with elements $\bar{N}_k \equiv \sqrt{W_k^{(l)}}|\tilde{N}^{(l)}(k)|$ and $\bar{X}_k^H \equiv \sqrt{W_k^{(l)}}|\tilde{X}^{(l)}(k)||\tilde{H}^{(l)}(k)|$. Hence

$$|\alpha^{(l)}| \leq \frac{<\bar{N}, \bar{X}^H>}{|\bar{N}||\bar{X}^H|} \leq 1.$$

Further, we define the log Mel-filter-bank energy (log-spectrum) vectors:

$$\mathbf{y} = \begin{bmatrix} \log|\tilde{Y}^{(1)}|^2 \\ \log|\tilde{Y}^{(2)}|^2 \\ \vdots \\ \log|\tilde{Y}^{(l)}|^2 \\ \vdots \\ \log|\tilde{Y}^{(L)}|^2 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} \log|\tilde{X}^{(1)}|^2 \\ \log|\tilde{X}^{(2)}|^2 \\ \vdots \\ \log|\tilde{X}^{(l)}|^2 \\ \vdots \\ \log|\tilde{X}^{(L)}|^2 \end{bmatrix},$$

$$\mathbf{n} = \begin{bmatrix} \log|\tilde{N}^{(1)}|^2 \\ \log|\tilde{N}^{(2)}|^2 \\ \vdots \\ \log|\tilde{N}^{(l)}|^2 \\ \vdots \\ \log|\tilde{N}^{(L)}|^2 \end{bmatrix}, \mathbf{h} = \begin{bmatrix} \log|\tilde{H}^{(1)}|^2 \\ \log|\tilde{H}^{(2)}|^2 \\ \vdots \\ \log|\tilde{H}^{(l)}|^2 \\ \vdots \\ \log|\tilde{H}^{(L)}|^2 \end{bmatrix}, \quad (7)$$

and define the vector of phase factors:

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha^{(1)} \\ \alpha^{(2)} \\ \vdots \\ \alpha^{(l)} \\ \vdots \\ \alpha^{(L)} \end{bmatrix}.$$

Then, we rewrite Eq. 5 as

$$\begin{aligned} e^{\mathbf{y}} &= e^{\mathbf{x}} \bullet e^{\mathbf{h}} + e^{\mathbf{n}} + 2\,\boldsymbol{\alpha} \bullet e^{\mathbf{x}/2} \bullet e^{\mathbf{h}/2} \bullet e^{\mathbf{n}/2} \\ &= e^{\mathbf{x}+\mathbf{h}} + e^{\mathbf{n}} + 2\,\boldsymbol{\alpha} \bullet e^{(\mathbf{x}+\mathbf{h}+\mathbf{n})/2}, \end{aligned} \quad (8)$$

where the $\bullet$ operation for two vectors denotes element-wise product, and each exponentiation of a vector above is also an element-wise operation. To obtain the log Mel-filter-bank energy for noisy speech, we apply the log operation on both sides of Eq. 8:

$$\begin{aligned}\mathbf{y} &= \log\left[e^{\mathbf{x+h}} \bullet (\mathbf{1} + e^{\mathbf{n-x-h}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{x+h+n}}{2}-\mathbf{x-h}})\right] \\ &= \mathbf{x} + \mathbf{h} + \log[\mathbf{1} + e^{\mathbf{n-x-h}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{n-x-h}}{2}}] \\ &\equiv \mathbf{y}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \boldsymbol{\alpha}). \end{aligned} \quad (9)$$

From Eq. 9, the phase factor (vector) $\boldsymbol{\alpha}$ can be solved as a function of the remaining variables:

$$\begin{aligned}\boldsymbol{\alpha} &= \frac{e^{\mathbf{y-x-h}} - e^{\mathbf{n-x-h}} - \mathbf{1}}{2e^{\frac{\mathbf{n-x-h}}{2}}} \\ &= 0.5(e^{\mathbf{y}-\frac{\mathbf{n+x+h}}{2}} - e^{\frac{\mathbf{n-x-h}}{2}} - e^{-\frac{\mathbf{n-x-h}}{2}}) \\ &\equiv \boldsymbol{\alpha}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \mathbf{y}) \end{aligned} \quad (10)$$

Eq.9 or Eq.10 constitutes the (deterministic) version of the phase-sensitive model for environmental distortion. One can be used for model adaptation and the other used for feature enhancement, which have been implemented in [21] and [10], respectively, after extending them into the probabilistic version which we describe below.

5.3. The phase-sensitive model of environmental distortion — Probabilistic version

We now use the nonlinear relationship among the phase factor $\boldsymbol{\alpha}$ and the log-domain signal quantities of $\mathbf{x}, \mathbf{n}, \mathbf{h}$, and $\mathbf{y}$, as derived above and shown in Eqs. 9 or 10, as the basis to develop a probabilistic phase-sensitive model for the acoustic environment. The outcome of a probabilistic model for the acoustic environment is explicit determination of the conditional probability, $p(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h})$, of noisy speech observations ($\mathbf{y}$) given all other variables $\mathbf{x}, \mathbf{n}$, and $\mathbf{h}$. This conditional probability is what is required in the Bayesian network model to specify the conditional dependency as denoted by each arrow in Fig. 1. This conditional probability is also required for deriving an optimal estimate of clean speech, which was carried out in the work of [10].

To determine the form of $p(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h})$, we first need to assume a form of the statistical distribution for the phase factor $\boldsymbol{\alpha} = \{\alpha^{(l)}, l = 1, 2, ..., L\}$. To accomplish this, we note that the angle θ_k between the complex variables of $N[k]$ and $(X[k]H[k])$ is uniformly distributed over $(-\pi, \pi)$. This amounts to the maximal degree of randomness in mixing speech and noise, and has been empirically observed to be correct.

Then, from the definition of $\alpha^{(l)}$ in Eq. 6, it can be shown that the phase factor $\alpha^{(l)}$ for each Mel-filter l can be approximated by a (weighted) sum of a number of independent, zero-mean random variables $\cos(\theta_k)$ distributed (non-uniformly but symmetrically) over $(-1, 1)$, where the total number of terms equals the number of DFT bins (with a non-zero gain) allocated to the Mel-filter. When the number of terms becomes large, as is typical for high-frequency filters, the central limit theorem postulates that $\alpha^{(l)}$ will be approximately Gaussian. Law of large numbers further postulates that the Gaussian will have the mean of zero since each term of $\cos(\theta_k)$ has the mean of zero.

Thus, the statistical distribution for the phase factor can be reasonably assumed to be a zero-mean Gaussian:

$$p(\alpha^{(l)}) = \mathcal{N}(\alpha^{(l)}; 0, \Sigma_\alpha^{(l)}),$$

where the filter-dependent variance $\Sigma_\alpha^{(l)}$ is estimated from a set of training data. Since noise and (channel-distorted) clean speech are mixed independently for each DFT bin, we can also reasonably assume that the different components of the phase factor $\boldsymbol{\alpha}$ are uncorrelated. Thus, we have the multivariate Gaussian distribution of

$$p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}; \mathbf{0}, \boldsymbol{\Sigma}_\alpha), \quad (11)$$

where $\boldsymbol{\Sigma}_\alpha$ is a diagonal covariance matrix.

Given $p(\boldsymbol{\alpha})$, we are now in a position to derive an appropriate form for $p_y(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h})$. To do so, we first fix the values of $\mathbf{x}, \mathbf{n}$, and $\mathbf{h}$, treating them as constants. We then view Eq. 9 as a (monotonic) nonlinear transformation from random variables $\boldsymbol{\alpha}$ to $\mathbf{y}$. Using the well-known result from probability theory on determining the PDF for functions of random variables, we have

$$p_y(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h}) = |J_{\boldsymbol{\alpha}}(\mathbf{y})|\, p_\alpha(\boldsymbol{\alpha}|\mathbf{x}, \mathbf{n}, \mathbf{h}), \quad (12)$$

where $J_{\boldsymbol{\alpha}}(\mathbf{y}) = \frac{1}{\frac{\partial \mathbf{y}}{\partial \boldsymbol{\alpha}}}$ is the Jacobian of the nonlinear transformation.

The diagonal elements of the Jacobian can be computed, using Eq. 9 and then using Eq. 8, by

$$\begin{aligned}\mathrm{diag}\left(\frac{\partial \mathbf{y}}{\partial \boldsymbol{\alpha}}\right) &= \frac{2e^{\frac{\mathbf{n-x-h}}{2}}}{\mathbf{1} + e^{\mathbf{n-x-h}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{n-x-h}}{2}}} \\ &= \frac{2e^{\frac{\mathbf{n+x+h}}{2}}}{e^{\mathbf{x+h}} + e^{\mathbf{n}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{n+x+h}}{2}}} \\ &= 2\, e^{\frac{\mathbf{n+x+h}}{2} - \mathbf{y}}. \end{aligned} \quad (13)$$

The determinant of the diagonal matrix of Eq. 13 is then the product of all the diagonal elements.

Also, the Gaussian assumption for $\boldsymbol{\alpha}$ gives

$$p(\boldsymbol{\alpha}|\mathbf{x}, \mathbf{n}, \mathbf{h}) = p\left[\boldsymbol{\alpha}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \mathbf{y})\right] = \mathcal{N}\left[\boldsymbol{\alpha}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \mathbf{y}); \mathbf{0}, \boldsymbol{\Sigma}_\alpha\right]. \quad (14)$$

Substituting Eqs. 13 and 14 into Eq. 12, we establish the following probabilistic model of the acoustic environment:

$$\begin{aligned} p_y(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h}) = \frac{1}{2}\ &| \mathrm{diag}\left(e^{\mathbf{y}-\frac{\mathbf{n+x+h}}{2}}\right) | \\ &\mathcal{N}\left[\frac{1}{2}(e^{\mathbf{y}-\frac{\mathbf{n+x+h}}{2}} - e^{\frac{\mathbf{n-x-h}}{2}} - e^{-\frac{\mathbf{n-x-h}}{2}}); \mathbf{0}, \boldsymbol{\Sigma}_\alpha\right] \end{aligned} \quad (15)$$

Because α is the inner product (proportional to cosine of the phase) between the Mel-filter vectors of noise and clean speech characterizing their phase relationship, a Gaussian distribution on it makes the environment model of Eq. 15 phase sensitive.

An alternative, simplified form of the model in Eq. 15 can be obtained using linearized Taylor series approximation.

5.4. Experiments and their analysis

A principled way of carrying out environment-robust ASR, in conjunction with handling other sources of variability in an integrative manner, is to directly use the probabilistic, phase-sensitive model of Eq. 15 as the conditional dependency from the o-layer (i.e., $\mathbf{x}$), n-layer, and h-variable to the y-layer in the multi-layer Bayesian network model of Fig. 1, and then to carry out probabilistic inference. The probabilistic inference for the (continuous-valued) o-layer variables gives algorithms for clean-speech feature enhancement. The probabilistic inference for the (discrete-valued) S-layer variables gives algorithms for phonetic or word recognition of distorted speech with the y-layer input.

All existing experiments reported in the literature using the phase-sensitive model, however, have used much simplified higher-layer models in Fig. 1. The work reported in [34, 21] made use of HMMs as the higher-layer model; I.e., removing the t-layer and z-layer in Fig. 1 altogether and making direct dependency from the S-layer to o-layer. The work reported in [10, 11] used an even simpler higher-layer model (Gaussian mixture model) by further removing the temporal conditional dependency in the S-layer of Fig. 1. In the latter work, the S-layer does not represent any phonetic information and hence the overall approach in [10, 11] can no longer be used for speech recognition directly but rather it is used for feature enhancement or making inference on the "hidden" o-layer variables. Then, the enhanced features are used in a separate, pre-trained HMM system to perform speech recognition.

5.4.1. Results on feature enhancement using the phase-sensitive model

As reported in [10], a diagnostic experiment was carried out to assess the role of phase asynchrony in feature enhancement for noise-robust ASR. To eliminate the factor of noise power estimation inaccuracy, phase-removed true noise power is used since in the Aurora2 task the true noise's waveforms are made readily available. Table 1 lists the percent accuracy in the Aurora2 standard task of digit recognition (as a function of the feature enhancement algorithm iterations using the phase-sensitive model; see the algorithm in [10]). Clean HMMs (simple backend) as provided by the Aurora2 task are used for recognizing enhanced features.

When the phase information is removed, how much does the performance suffer? To examine this issue, several spectral subtraction methods are used where the same phase-removed true noise features are used as in Table 1. After careful tuning of the spectral subtraction parameter of the floor value, the best accuracy is 96% (see detailed results in Table 2), significantly below the accuracy of 98% obtained with the use of the phase-sensitive model.

Table 1. Percent accurate digit recognition rate for the Aurora2 task as a function of the feature enhancement algorithm iteration number using the phase-sensitive model. Phase-removed true noise features (noise power spectra) are used in this diagnostic experiment as the n-layer variables.

Itrs	1	2	4	7	12
SetA	94.12	96.75	97.96	98.11	98.12
SetB	94.80	97.29	98.10	98.48	98.55
SetC	91.00	94.50	96.50	97.86	98.00
Ave.	93.77	96.52	97.72	98.21	**98.27**

Table 2. Performance (percent accurate) for the Aurora2 task using four versions of spectral subtraction (SS) with the same phase-removed true noise features as in Table 1.

Floor	e^{-20}	e^{-10}	e^{-5}	e^{-3}	e^{-2}
SS1	93.57	94.26	**95.90**	92.18	90.00
SS2	12.50	44.00	65.46	88.69	84.44
SS3	88.52	89.26	93.19	90.75	88.00
SS4	10.00	42.50	63.08	87.41	84.26

However, instead of using true noise power, when the estimated noise power is used (with the algorithm for noise power estimation described in [10]), improvement of recognition accuracy from using the phase-insensitive model to the phase-sensitive model becomes much smaller, from 84.80% only to 85.74%; see detailed results in Table 3).

What may be the reason for the drastic difference between the performance improvements (from the phase-insensitive to phase-sensitive models) with and without noise estimation errors? Let us examine Eq. 5. It is clear that the third, phase-related term in Eq. 5 and the second, noise-power term are added to contribute to the power of noisy speech. If the estimation error in the second, noise-power term is comparable to the entire third term, then the addition of the third term would not be very meaningful in accounting for the power of noisy speech. This is the most likely explanation for the huge performance improvement when true noise power is used (Tables 1 and 2) and relatively mild improvement when noise power estimation contains errors (Table 3). The analysis above shows

Table 3. Right column: percent accurate digit recognition rates for the Aurora2 task using noise estimation and phase-sensitive feature enhancement algorithms, both described in [10]. Left column: The baseline results obtained with the phase-insensitive model.

	Baseline (no phase)	Enhanced (with phase)
SetA	85.66	86.39
SetB	86.15	86.30
SetC	80.40	83.35
Ave.	**84.80**	**85.74**

the critical role of noise power estimation in enabling the effectiveness of using the phase-sensitive model of environmental distortion.

5.4.2. Results on HMM adaptation using the phase-sensitive model

In the more recent work of [21], the phase-sensitive model in the cepstral domain and its Taylor series approximation are used to adapt/estimate the HMM parameters and the same phase-sensitive model is used to estimate the noise power. The joint estimation of both the HMM and noise parameters using the consistent distortion model gives better quality of the noise estimate than that by the noise estimation algorithm presented in [10] producing the results shown in Table 3. As a consequence, the performance improvement in the same Aurora2 task from using the phase-insensitive to phase-sensitive models is much greater: Digit recognition rate of $\mathbf{91.70}\%$ (using the phase-insensitive model of environmental distortion to adapt HMM parameters [20]) is improved to $\mathbf{93.32}\%$ after the use of the phase-sensitive model, giving 19.5% relative error rate reduction.

In the earlier work of [34], the same phase-sensitive model was exploited to adapt HMM parameters but with very different approximation. Also, while crucial in understanding the effectiveness of using the phase-sensitive model, no noise estimation technique was described. A different evaluation task, Aurora4 dictation, was used in evaluating the adaptation method, and 3.5% relative error reduction was reported.

6. SUMMARY AND CONCLUSION

Acoustic modeling and robust speech recognition have been and are continuing to be active research areas. In this paper, I argue the case for the important roles that high-fidelity acoustic models can play in robust speech recognition. High-fidelity (vs. "lower-fidelity") modeling refers to the use of a richer set of useful, albeit incomplete, knowledge in constructing probabilistic models of the speech process for the purpose of speech class discrimination (with or without an intermediate process of speech feature generation). Robustness refers to the maintenance of high performance in speech recognition against pervasive and inherent variability including both speaker and environmental factors. Two detailed case studies are presented in this paper, relating to each of these two factors. In the first case study, the high-fidelity acoustic model constructed using a multi-layer dynamic Bayesian network is described that embeds the knowledge of articulatory-like constraints that govern the dynamic pattern in the formant or VTR movement from one speech unit to the another throughout the speech utterance. This model is shown in computer simulation to account for realistic formant reduction and the consequent increase in "static" phonetic confusability, both being difficult to produce in the acoustic models based on HMMs. In the second case study, a high-fidelity phase-sensitive model of environmental distortion is derived that embeds the knowledge of phase asynchrony between clean speech and the mixing noise, commonly ignored in the standard techniques. Experimental results are reviewed from the literature that demonstrate the effectiveness of the model in noise-robust speech recognition over the commonly used phase-insensitive model, especially when the estimate of noise power is reasonably accurate. In particular, when noise power estimation contains no errors in our diagnostic experiments, the results showed over 50% error reduction moving from the use of the phase-insensitive to phase-sensitive models in speech feature enhancement prior to speech recognition.

Future directions of research in acoustic modeling include the incorporation of more structured and beneficial knowledge about the nature of speech variability into probabilistic models, and the development of more effective algorithms for learning and decision/decoding using such higher-fidelity models that make use of the more advanced knowledge. What kind of knowledge is likely to be most beneficial? The answer may be gleaned from the recent MINDS report [3], in which a number of speech recognition/understanding areas deemed especially fertile for future research are identified. Relevant to acoustic modeling and robust speech recognition are the following rich areas for future research: 1) Advanced acoustic models and architectures that can handle "everyday audio", with the focus on robustness of the speech system for meaningful acoustic environments as diverse as meeting room presentations to unstructured conversations, and on rapid adaptation to changing acoustic conditions in multiple dimensions, even simultaneously. 2) Adaptation and self-learning in speech recognition system, with the focus on learning from poorly labeled or even un-annotated data, and on generalization to enable recognizers to operate effectively in novel circumstances (e.g., different tasks, environments, and languages). 3) Cognition-derived speech models and algorithms, with the focus on understanding and emulating relevant human capabilities in speech processing and on incorporating these strategies into automatic speech systems. Of particular interest are how signifi-

cant cortical information processing capabilities beyond signal processing are achieved and how one can leverage that knowledge in our automatic algorithms and systems. 4) Effective representation and utilization of knowledge sources drawing from fundamental science of human speech perception and production, with the focus on the essential properties that underlie auditory masking and attention, on emulating human capabilities to rapidly adapt to non-native accents, and on the temporal span over which signal signals are represented, produced, and modeled. All of the above research directions require the construction of robust acoustic models, as well as related algorithms, that are of much higher "fidelity" than the ones discussed in the two case studies presented in this paper.

7. REFERENCES

[1] A. Acero. Acoustical and Environmental Robustness in Automatic Speech Recognition. Kluwer Academic Publishers, 1993.

[2] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," Proc. ICSLP, Vol.3, pp. 869-872, 2000.

[3] J. Baker, L. Deng, S. Khudanpur, C. Lee, J. Glass, and N. Morgan. "Historical Development and Future Directions in Speech Recognition and Understanding," MINDS Report of the Speech Understanding Working Group, NIST, 2006-2007. http://www.itl.nist.gov/iad/894.02/MINDS/FINAL/speech.web.pdf

[4] J. Bilmes and C. Bartels. "Graphical model architectures for speech recognition," IEEE Signal Processing Magazine, vol. 22, no. 5, September 2005, pp. 89-100.

[5] W. Chou and F. Juang (eds.) Pattern Recognition in Speech and Language Processing, CRC Press, 2003.

[6] L. Deng. Dynamic Speech Models — Theory, Algorithm, and Application, Morgan & Claypool Publishers, LaPorte CO, USA, 2006.

[7] L. Deng and D. Sun. "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acoust. Soc. Am.*, Vol. 95, 1994, pp. 2702–2719.

[8] L. Deng, D. Yu, and A. Acero. "A bi-directional target-filtering model of speech coarticulation and reduction: Two-stage implementation for phonetic recognition," *IEEE Trans. Speech & Audio Processing*, Vol. 14, No. 1, January 2006, pp. 256-265.

[9] L. Deng, D. Yu, and A. Acero. "Structured speech modeling," IEEE Transactions on Audio, Speech and Language Processing (Special Issue on Rich Transcription), Vol. 14, No. 5, Sept 2006, pp. 1492-1504.

[10] L. Deng, J. Droppo, and A. Acero. "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," IEEE Trans. on Speech and Audio Processing. Vol.12 (2), Mar 2004. pp. 133-143.

[11] J. Droppo, A. Acero, and L. Deng. "A nonlinear observation model for removing noise from corrupted speech log Mel-spectral energies, Proc. ICSLP, Denver, Colorado, Sept, 2002.

[12] V. Goel and W. Byrne. "Minimum Bayes-risk automatic speech recognition," Computer Speech and Language, 2000, Vol. 14(2), pp. 115135.

[13] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," IEEE Trans. Speech and Audio Proc., Vol. 13, No. 5, pp. 975-983, 2005.

[14] X. He, L. Deng, and W. Chou. "Discriminative learning in sequential pattern recognition — A unifying review ," IEEE Signal Processing Magazine, 2008, to appear.

[15] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," Proc. ISCA ITRW ASR, 2000.

[16] L. Rabiner and B.-H. Juang. Fundamentals of Speech Recognition. Prentice Hall, 1993.

[17] J.- C. Junqua. Robust Speech Recognition in Embedded Systems and PC Application, Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[18] C.-H. Lee and Q. Huo. "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," Proc. of the IEEE, Vol. 88, No. 8, Aug. 2000, pp. 1241-1269.

[19] C. -H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," Speech Communication, Vol. 25, pp. 29-47, 1998.

[20] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions," Proc. IEEE ASRU, 2007, to appear.

[21] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero. "HMM adaptation using a phase-sensitive acoustic distortion model for environment-robust speech recognition," Submitted to ICASSP 2008.

[22] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," Proc. ICASSP, Vol. IV, pp. 389-392, 2007.

[23] E. McDermott, T. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large vocabulary speech recognition using minimum classification error," IEEE Trans. Speech and Audio Processing, Vol. 15. No. 1, 2007, pp. 203-223.

[24] P. Moreno. Speech Recognition in Noisy Environments. PhD. Thesis, Carnegie Mellon University, 1996.

[25] N. Morgan, Q. Zhu, A. Stolcke, etc. "Pushing the envelope — Aside," *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005, pp. 81-88.

[26] H. Ney. "Bayes decision rule, classification error, and training criteria," Lecture Notes, RWTH Aachen - University of Technology, 2006.

[27] H. Nock and S. Young. "Loosely coupled HMMs for ASR: A preliminary study," Technical Report TR386, Cambridge University, 2000.

[28] M. Ostendorf. "Moving beyond the beads-on-a-string model of speech" *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, December 1999, 9-83, Keystone, CO. pp. 79-83.

[29] M. Ostendorf, V. Digalakis, and J. Rohlicek. "From HMMs to segment models: A unified view of stochastic modeling for speech recognition" *IEEE Trans. Speech & Audio Processing*, Vol. 4, 1996, pp. 360-378.

[30] F. Pereira. "Linear models for structure prediction," Proc. Interspeech, Lisbon, 2005, pp. 717-720.

[31] M. Pitermann. "Effect of speaking rate and contrastive stress on formant dynamics and vowel perception," *J. Acoust. Soc. Am.*, Vol. 107, 2000, pp. 3425-3437.

[32] D. Povey. "Discriminative Training for large Vocabulary Speech Recognition," Ph.D. dissertation, Cambridge University, Cambridge, UK, 2004.

[33] R. Schlter, W. Macherey, B. Muller, and H. Ney. "Comparison of discriminative training criteria and optimization methods for speech recognition," Speech Communication, Vol. 34, 2001, pp. 287-310.

[34] V. Stouten, H. Van hamme, P. Wambacq. "Effect of phase-sensitive environment model and higher order VTS on noisy speech feature enhancement," Proc. ICASSP, 2005, pp. 433- 436.

[35] J. Sun and L. Deng. "An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition," *J. Acoust. Soc. Am.*, Vol. 111, No. 2, February 2002, pp.1086-1101.

[36] O. Viikki (Ed.). *Speech Communication* (Special Issue on Noise Robust ASR), Vol. 34, April 2001.

[37] Q. Zhu and A. Alwan. "The effect of additive noise on speech amplitude spectra: A quantitative analysis," *IEEE Signal Proc. Letters*, Vol. 9(9), Sept. 2002, pp. 275-277.