THE BLAME GAME IN MEETING ROOM ASR: AN ANALYSIS OF FEATURE VERSUS MODEL ERRORS IN NOISY AND MISMATCHED CONDITIONS

Sree Hari Krishnan Parthasarathi[1], Shuo-Yiin Chang[1,2], Jordan Cohen[1], Nelson Morgan[1,2], Steven Wegmann[1]

[1]International Computer Science Institute, Berkeley, USA.
[2]UC Berkeley, Berkeley, USA.

{sparta,shuoyiin,morgan,swegmann}@icsi.berkeley.edu, jordan.jordan.cohen@gmail.com

ABSTRACT

Given a test waveform, state-of-the-art ASR systems extract a sequence of MFCC features and decode them with a set of trained HMMs. When this test data is clean, and it matches the condition used for training the models, then there are few errors. While it is known that ASR systems are brittle in noisy or mismatched conditions, there has been little work in quantitatively attributing the errors to features or to models. This paper attributes the sources of these errors in three conditions: (a) matched near-field, (b) matched far-field, and a (c) mismatched condition. We undertake a series of diagnostic analyses employing the bootstrap method to probe a meeting room ASR system. Results show that when the conditions are matched (even if they are far-field), the model errors dominate; however, in mismatched conditions features are neither invariant nor separable and this causes as many errors as the model does.

Index Terms— Features, acoustic conditions, hidden Markov models, speech recognition.

1. INTRODUCTION

In this paper we will present a quantitative analysis that partially answers the question "why is automatic speech recognition so brittle?" One of the major contributing factors to this brittleness is the remarkable inability of the standard acoustic model, the hidden Markov model (HMM), to accurately model speech test data that differs in character from the data that was used for its training. While there has long been speculation about the root causes of this brittleness, ranging from the over-fitting of the acoustic model to its training data to the lack of invariance of the standard front-end (mel-frequency cepstral coefficients (MFCCs)), there is surprisingly little quantitative evidence available to back up one claim over another. Furthermore, the research aimed at improving HMM-based speech recognition accuracy has largely ignored questions concerning understanding or quantifying the underlying causes of recognition errors with notable exceptions being [1, 2]. Instead, improvements–many of which are reviewed in [3, 4, 5, 6, 7]–to the front-end and the acoustic models have largely proceeded by trial and error. The research that we will describe is a continuation of the research described in [8, 9] that used simulation and a novel sampling process to quantify the effects that the two major HMM assumptions have on recognition accuracy. In this previous work, we analyzed recognition performance on tasks[1] where the properties of the training and test acoustic data were not challenging and were homogeneous, or matched, across the training and test sets. In this paper, however, we will be analyzing recognition performance using the ICSI meeting corpus [10] where the acoustic data are more challenging and we are able to exploit properties of this corpus to compare recognition performance when the training and test data acoustics are matched or mismatched.

More specifically, we use the parallel recordings using near and far-field microphones in the ICSI meeting corpus [10] to construct three sets of related recognition tasks: (a) matched near-field acoustic model training and recognition test data; (b) matched far-field acoustic model training and recognition test data; (c) mismatched near-field acoustic model training data and far-field recognition test data. The results of our analysis in the matched cases (a) and (b) are identical to what we found in [9], namely that 1) long range statistical dependence that is present in speech data and at variance with the HMM's conditional independence assumption is the single largest source of recognition errors and that 2) that MFCCs are essentially separable under this model. However, the results of our analysis is quite different in the mismatched case (c): here we demonstrate that the lack of invariance that MFCCs exhibit to the transformation between the near and far-field acoustics is a major source of recognition errors, approximately equal to the number of recognition errors caused by statistical dependence in the data. Together these two sources of errors dominate in the mismatched case and are the major cause of brittleness in automatic speech recognition.

The experimental methodology of this paper is quite involved, and it drives the overall structure. Section 2 outlines the preparation of the data in terms of partitioning and time alignments. Model building and resampling are presented in Sections 3 and 4. The main diagnostic experiments, followed by an interlude on adaptation, are reported in Section 5. Concluding remarks are presented in Section 6.

2. DATASETS

We are using a dataset of spontaneous meeting speech recorded at ICSI [10] where each spoken utterance was captured using near-field (NF) and far-field[2] (FF) microphones. Our training set is based on the meeting data used for adaptation in the SRI-ICSI meeting recognition system [11]. For the test set we used the ICSI meetings drawn from the NIST RT eval sets [12, 13, 14]; this was done to control the variability in the data for the resampling experiments.

The remainder of this section discusses the creation of the parallel NF and FF corpora for this paper. First we describe how we estimate and remove a variable length time delay that exists between the corresponding NF and FF utterances, so that each training and test utterance has two parallel versions–NF and FF–that line up at the MFCC frame level. Next we discuss how we partition these parallel NF and FF corpora data into training and test sets.

[1]Based on the Wall Street Journal and Switchboard corpora.

[2]We used the "single distant microphone" recordings for the far-field data.

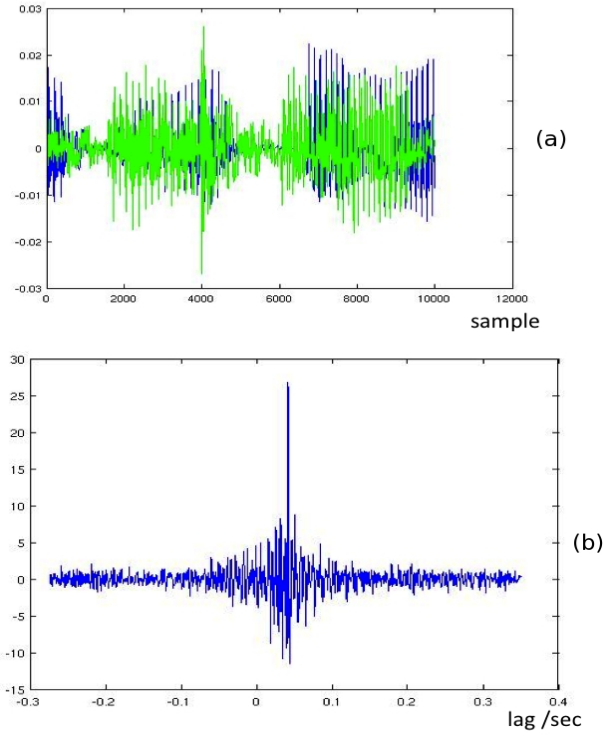

Fig. 1. Time alignment: (a) NF (blue) and FF (green) signals (b) Cross-correlation between the signals.

2.1. Time-aligning the corpora

In order to synchronize the NF and FF recordings, we must deal with a time delay, or skew, that exists between the two recordings. These time delays arise from two factors: (1) different physical distances between the speakers and the microphones, and (2) systematic delays introduced by the recording software. The latter factor appears to dominate the skew between the NF and FF recordings. Fixed delays were introduced when the channels were initialized at the start of a recording. Since this systematic delay dominates the skew, the NF recordings have a time delay relative to the FF recordings. Fig 1(a) illustrates an utterance captured by the FF microphone that is advanced in time in comparison to the same utterance captured by the NF microphone.

Time delay is more evident in the cross-correlation between the NF and FF signals, as shown in Fig 1(b). The delay could be estimated by searching for a peak in the cross-correlation sequence. In Fig 1(b) the peak is at a lag of 41.88 ms (670 samples at 16 kHz). However, this detection could be difficult because of the recording quality and noise. To guarantee a more precise detection, we divide each utterance into overlapping windows, where the window size is a third of the utterance length and the step size for successive windows is a tenth of the utterance length. For each step, the cross-correlation sequence is calculated and a delay is estimated. If the variation between the estimated delays in the windows for a given utterance is too large, then the estimated delay is regarded as unreliable and the utterance is discarded. Approximately 30% of the utterances were discarded because of these unreliable delay estimates. The delays between NF and FF channels for the reliable data ranged from 12.5 ms to 61.25 ms. This was implemented using the Skewview tool [15]. A more detailed discussion of the time delay can be found in [16].

2.2. Data partitions

Because of the parallel nature of the NF and FF corpora, the data partitions are identical. For simplicity, we describe the NF partitioning. The training set had a dominant speaker accounting for nearly a quarter; clearly this would skew the data generated by the resampling process. On the other hand, perfect speaker balancing cannot be achieved given that this is a corpus of spontaneous speech. There is, therefore, a trade-off between "the amount of data" and an "egalitarian distribution of speakers". The resulting NF training and test sets consists of about 20 hours and 1 hour respectively and their statistics are reported in Table 1.

Table 1. *Training and test statistics for NF and FF.*

Dataset	Speakers	Utterances	Time
Training	26	23729	20.4 (hrs)
Test	18	1063	57.9 (mins)

3. MODELS AND EXPERIMENTAL SETUP

We use version 3.4 of the HTK toolkit [17] for the front-end, acoustic model training, and decoding. In particular, we use the standard HTK front-end to produce a 39 dimensional feature vector every 10 ms: 13 Mel-cepstral coefficients, including energy, plus their first and second differences. The cepstral coefficients are mean-normalized at the utterance level. We use HDecode for decoding with a wide search beam (300) to avoid search errors. To evaluate recognition accuracy the reference and the decoded utterances are text normalized before the NIST tool sclite is used to obtain word error rate (WER). The remainder of this section discusses the recognition acoustic models, dictionary, and language model.

3.1. Near-field acoustic models

The NF acoustic models use cross-word triphones and are estimated using maximum likelihood. Except for silence, each triphone is modeled using a three-state HMM with a discrete linear transition structure that prevents skipping. The output distribution for each HMM state is a single, multivariate Gaussian with diagonal covariance. While signicantly better performance can be achieved with mixtures of more components, the simplicity of a single component is preferable for our analysis; it also highlights the performance differences between our experiments. Maximum likelihood training roughly follows the HTK tutorial: monophone models are estimated from a "flat start", duplicated to form triphone models, clustered to 2500 states and re-estimated.

3.2. Far-field acoustic models: via single-pass retraining

Instead of building the FF acoustic models from a flat start, we exploit the parallel nature of the NF and FF training sets to build the FF models using *single-pass retraining* from the final NF models and the FF data. Single-pass retraining is a form of EM, which is supported by HTK, where, in our case, the E-step is performed using the NF models and data, while the M-step and model updates use the FF data. We only update the means and variances of the FF models, so the result is a parallel set of NF and FF acoustic models that share the same state-tying but the (unknown) transformation between the NF and FF means and variances is determined by the frame-level transformation between the parallel NF and FF acoustic data.

3.3. Dictionary and language models

Since we are using relatively simple acoustic models– single mixture component per state and 2500 tied states–and that the recognition task is much more complex compared to [8, 9], we use a powerful language model (LM) to keep the error rate manageable. In fact, our initial experiments using a weaker LM derived from the training set resulted in WERs as high as 64% in the matched NF condition.

We use a LM [18] that was trained at SRI by interpolating a number of source LMs; these consisted of webtext and the transcripts of the following corpora: Switchboard, meetings (CMU, ICSI, and NIST), Fisher, Hub4-LM96, and TDT4. We then removed words not in the training dictionary from the trigram LM, and renormalized it. The perplexity of this meeting room LM is around 70 on our test set. To avoid out-of-vocabulary issues, all test utterances containing a word not present in the LM are removed. To be compatible with the SRI LM, we use the SRI pronunciation dictionary; it uses two extra phones in comparison with the CMU phone set–"puh" and "pum"–for hesitations.

4. SIMULATION AND RESAMPLING METHODOLOGY

We use simulation and a novel sampling process to generate pseudo test data that deviate from the major HMM assumptions in a controlled fashion. The novel sampling process, called resampling, was adapted from Bradley Efron's work on the bootstrap [19] in [8, 9]. These processes allow us to generate pseudo data that, at one extreme, agree with all of the model's assumptions, and at the another extreme, deviate from the model in exactly the way real data do. In between, we can precisely control the degree of data/model mismatch. By measuring recognition performance on this pseudo test data, we are able to quantify the effect of this controlled data/model mismatch on recognition accuracy.

4.1. The simulation and resampling process

Our methodology allows six levels of simulation and resampling: (a) simulation (b) frame resampling (c) state resampling (d) phone resampling (e) word resampling (f) original test utterance.

Simulation: We follow the full generative process assumed by HMMs. The simulated data, therefore, matches all the assumptions of the model. These assumptions are: (a) the sequence of states are hidden and are constrained to follow a Markov chain (b) the features are independent conditioned on the states (c) the output distributions are stationary and can be modeled using a single Gaussian.

To generate the test data by simulation, we start with the test transcriptions, and look up each word in the pronunciation dictionary to create phone transcriptions. We then use the state transitions and the output distribution associated with the states belonging to the triphones to generate the data. Note however that the delta and acceleration features are also generated.

Frame resampling: In this case, we do not use the full generative process. Nevertheless, we create data that respects the independence assumptions at different levels. To generate the data in this fashion the following process is performed: (a) we use the training model is used to perform forced alignment on the training utterances, so that each speech frame is annotated with its most likely generating state. (b) We walk through this alignment, filling an urn for each state with its representative frames; at the end of this process, each urn is populated with frames representing its empirical distribution. (c) To generate resampled data, we use the model to create a forced alignment of the test data, and then sample a frame (at random, with replacement) from the appropriate urn for each frame position; these resampled frames are concatenated. With this frame-level resampling, the pseudo test data is exactly the same length as the original, and has the same underlying alignment, but the frames are now conditionally independent (given the state).

State, phone, and word resampling: By placing entire state sequences of frames in the urns, and then resampling (again, concatenating samples), we end up with pseudo test data with dependence among frames within state regions, but independence across state boundaries (note that resampling units larger than single frames produces pseudo test data that may be a different length from the original). We can further extend this idea to phones and to words; in all cases, the urn labels include the full triphone context.

4.2. Enforcing common alignment for NF and FF

In the previous sections, we described the methods used to ensure that the datasets and the models are completely parallel in the near-field and the far-field cases. This was done so that the errors in the mismatched case can be attributed solely to either the features or the models. However, one more variability remains, and that is in the resampling process.

The method of resampling creates an alignment of the training dataset using the recognition model; it then uses the alignments to fill urns that are in turn used to create the pseudo test utterances. The differences in the alignments created by the near-field and the far-field model will lead to the creation of pseudo test sets that are not parallel, leading to the near-field model trying to compensate, in addition, for a mismatched alignment. In order to minimize this effect, we create alignments using the near-field model on the near-field data, and use this alignment to generate pseudo, far-field test data (for the mismatched case).

5. RESULTS AND DISCUSSION

Near-field and far-field test data are created by simulation, resampling frames, states, phonemes, and words; then the corresponding recognition models are used for decoding. Each resampling experiment is repeated five times and the results are shown in the Table 2. In the matched NF experiments, NF models are used to recognize NF test data, while the matched FF experiments use FF models and FF test data. In the mismatched experiments, NF models are used to recognize FF test data. Listed in the table for the matched and the mismatched cases are the word error rate (WER), standard error (SE), and the relative increase in WER from previous level of simulation/resampling (the next highest row). The standard errors range from 0.03 (simulation in the NF case) to 0.45 (word resampling in the FF case), so all the WER differences between matched and mismatched conditions are significant. Note that the WERs on the test data increase as we move from NF (44.7%) to FF (71.4%), and then to the mismatched conditions (84.7%); this indicates the difficulty of the tasks.

5.1. Analysis of matched near-field results

It is remarkable to see that the WER for simulation and frame resampling is negligibly small in meeting room data, albeit with near-field microphones; for these cases all assumptions made by the model are satisfied by the data. When this is the case, the WER obtained by the system must be similar to human performance. The largest increase in WER is observed when we move from frame resampling to state resampling – a little more than a four-fold increase in errors.

Table 2. *Results for the matched (near-field and far-field) and the mismatched cases. For each of these cases, the word error rate (WER), standard error (SE), and the increase (%) in WER obtained over the next higher level of resampling are listed.*

Resampling	Near-field			Far-field			Mismatched		
	WER (%)	SE	ΔWER (%)	WER (%)	SE	ΔWER (%)	WER	SE	ΔWER (%)
Sim	1.4	0.03	-	1.8	0.03	-	43.0	0.23	-
Frame	1.9	0.05	31	3.4	0.02	88	59.9	0.26	39
State	9.6	0.17	416	23.2	0.2	580	75.8	0.27	27
Phone	21.4	0.21	123	45.5	0.41	96	80.6	0.29	6
Word	37.6	0.28	75	63.5	0.45	40	80.6	0.15	0
Original	44.7		19	71.4		12	84.7		5

Another large increase in WER (123%) occurs when we move down to phone resampling. As dependence is introduced (going down the rows), we start observing larger WER. These results are consistent with what we observed in [9] on the WSJ and Switchboard corpora, both of which which also had matched training and test conditions.

5.2. Analysis of matched far-field results

Although the WER is consistently worse for the FF results than the NF results, they are consistent with what we observe in the NF experiments and in [9]. However, it is striking how small the WER for simulation (1.8%) is when we consider how large the WERs are on real FF data (71.4%). This shows that, when the training and test conditions are matched, and the model assumptions implicit in HMM's are met, MFCC features are essentially separable even for the more challenging FF meeting data.

5.3. Analysis of the mismatched case

The results in the mismatched case are in stark contrast to those obtained for the matched cases. The WER for simulation is much higher at 43%, which indicates that MFCCs are not separable in this mismatched case. While the errors due to statistical dependence–the WER from the state resampling to the original data–are considerable (from 59.9% to 84.7%), they are no longer the dominant cause of recognition errors.

To better understand the mismatched simulation result, we compare it to the matched, NF simulation result. In both cases we use NF models to recognize simulated data: in the matched case this data is simulated by the NF models, while in the mismatched case this data is simulated from the FF models. Because we used single-pass retraining (Section 3.2) to create the FF models from the NF models, the unknown transformation between the NF and FF means and variances is inherited from the unknown transformation between the parallel NF and FF training utterances. Thus the transformation between the test utterances simulated from the NF and FF models is derived from the transformation between the NF and FF models, and it is related to, but much simpler than, the transformation between the parallel NF and FF training data. The NF models have a low WER on the simulated NF test data (1.4%), but they have a high WER (43%) on the simulated FF data which is transformed simulated NF data. If the features (MFCCs) were invariant to this transformation, then the WERs would be similar. However, since the WERs are very different, the features cannot be invariant, and the large difference in WERs is due to this lack of invariance.

5.4. Adaptation

A standard approach to mitigating recognition errors due to mismatched conditions is to perform unsupervised MLLR [20], a form of linear mean adaptation. Since the large difference between the matched NF and mismatched simulation and results is due to the lack of invariance of MFCCs to a (presumably) non-linear transformation between the NF and FF data, it is natural to try to compensate for this using MLLR. We treat the one hour of simulated test data as belonging to one speaker, and use the recognition hypotheses to generate the adaptation transforms for the NF models. We do two passes of adaptation: in the first pass a global adaptation is performed, while the second pass uses a regression class tree. We experimented with up to 16 regression classes in the second pass, but we found that 3 classes were optimal. In this case the simulation WER improves from 43.0% to 15.4%. While this is a large improvement, the adapted WER, 15.4%, is still much higher than the 1.4% WER on simulated NF data.

6. CONCLUSIONS

By exploiting the method of resampling, we constructed a series of pseudo datasets from near-field and far-field meeting room datasets, that at one end satisfied the HMM model assumptions, while at the other end deviated from the model in the way real data did. Using these datasets we probed the standard HMM/GMM framework for automatic speech recognition. Experiments show that when the conditions are matched (even if they are far-field), the model errors dominate; however, in mismatched conditions features are neither invariant nor separable, and contribute as much to the total errors as does the model. We then studied unsupervised MLLR adaptation as a means to compensate for this issue in the model space; while this approach mitigates the errors, the conclusions about the lack of invariance of the MFCC features in mismatched conditions still holds true. As part of future work, this study paves way for principled investigations into other spectro-temporal representations (say Gabor [21]).

7. ACKNOWLEDGMENTS

The authors would like to thank Adam Janin, Dan Ellis, and Andreas Stolcke for their help while setting up the dataset. Research funded in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the IARPA, the ODNI or the U.S. Government.

8. REFERENCES

[1] E Eide, H Gish, P Jeanrenaud, and A Mielke, "Understanding and improving speech recognition performance through the use of diagnostic tools," in *in Proc. ICASSP*, 1995.

[2] Lin Chase, *Error-Responsive Feedback Mechanisms for Speech Recognizers*, Ph.D. thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 1997.

[3] George Saon and Jen-Tzung Chien, "Large-vocabulary continuous speech recognition systems: A look at some recent advances," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 18 –33, nov. 2012.

[4] G. Heigold, H. Ney, R. Schluter, and S. Wiesler, "Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 58 –69, nov. 2012.

[5] M. Gales, S. Watanabe, and E. Fosler-Lussier, "Structured discriminative models for speech recognition: An overview," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 70 –81, nov. 2012.

[6] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 114 –126, nov. 2012.

[7] R.M. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 34 –43, nov. 2012.

[8] S. Wegmann and L. Gillick, "Why has (reasonably accurate) automatic speech recognition been so hard to achieve?," `arXiv:1003.0206 [cs.CL]`, 2010.

[9] D. Gillick, L. Gillick, and S. Wegmann, "Dont Multiply Lightly: Quantifying Problems with the Acoustic Model Assumptions in Speech Recognition," in *Proceedings of ASRU*. 2011, pp. 71–76, IEEE.

[10] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[11] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System," in *Proceedings of the Second International Workshop on Classification of Events, Activities, and Relationships (CLEAR 2007) and the Fifth Rich Transcription 2007 Meeting Recognition (RT 2007)*, 2007.

[12] "Rt-2002 evaluation plan," `http://www.itl.nist.gov/iad/mig/tests/rt/2002/docs/rt02_eval_plan_v3.pdf`.

[13] "Rt-04s evaluation data documentation," `http://www.itl.nist.gov/iad/mig/tests/rt/2004-spring/eval/docs.html`.

[14] "Rich transcription spring 2005 evaluation," `http://www.itl.nist.gov/iad/mig/tests/rt/2005-spring/index.html`.

[15] Dan Ellis, "Skewview tool," `http://labrosa.ee.columbia.edu/projects/skewview/`.

[16] "ICSI Meeting Alignments," `http://www1.icsi.berkeley.edu/~shuoyiin/research/meetingskew/chanskew.html`.

[17] S.J. Young, G. Evermann, MJF Gales, D. Kershaw, G. Moore, JJ Odell, DG Ollason, D. Povey, V. Valtchev, and PC Woodland, *The HTK book version 3.4*, 2006.

[18] O. Cetin and A. Stolcke, "Language modeling in the ICSI-SRI Spring 2005 meeting speech recognition evaluation system," Tech. Rep., International Computer Science Institute, 2005.

[19] B. Efron, "Bootstrap methods: another look at the jackknife," *Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[20] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Comput. Speech. Lang*, , no. 9, 1995.

[21] N Morgan, Q Zhu, A Stolcke, K Snmez, S Sivadas, T Shinozaki, M Ostendorf, P Jain, H Hermansky, D Gelbart, D Ellis, G Doddington, B Chen, etin, Herv Bourlard, and M Athineos, "Pushing the envelope aside: Beyond the spectral envelope as the fundamental representation for speech recognition," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 81 –88, 2005.