Auditory-model based robust feature selection for speech recognition

Christos Koniaris, Marcin Kuropatwinski, W. Bastiaan Kleijn

Sound and Image Processing Laboratory

School of Electrical Engineering

KTH - Royal Institute of Technology

Osquldas väg 10,

SE-100 44

Stockholm,

Sweden

chris.koniaris@ee.kth.se,

markurop@pg.gda.pl,

bastiaan.kleijn@ee.kth.se

(Dated: December 10, 2009)

Abstract

It is shown that robust dimension-reduction of a feature set for speech recognition can be based on a model of the human auditory system. Whereas conventional methods optimize classification performance, the new method exploits knowledge implicit in the auditory periphery, inheriting its robustness. Features are selected to maximize the similarity of the Euclidian geometry of the feature domain and the perceptual domain. Recognition experiments using MFCCs confirm the effectiveness of the approach, which does not require labeled training data. For noisy data the method outperforms commonly used discriminant-analysis based dimension-reduction methods that rely on labeling. The experiments indicate that selecting MFCCs in their natural order results in subsets with good performance.

PACS numbers: 43.71.-k,43.72.Qr,43.72.Ne

1. Introduction

The extraction of acoustic features is an essential component of automatic speech recognition (ASR). It enables the classification of speech signals at a reasonable computational complexity based on training with speech databases of a practical size. However, the data processing inequality implies that the extraction of acoustic features from a speech signal at best preserves information relevant for phone discrimination. Thus, careful selection of the acoustic features is essential.

The existing approach for selecting features from a larger set of candidate features is based on direct optimization of classification performance, using labeled training databases. Many algorithms have been developed to select features for classification[1–4]. In ASR it is common to use *dimension-reduction* procedures[4–6], a more general paradigm where input features are combined into a new set of lower cardinality. In general, existing feature-selection and dimension-reduction methods require classified training data. For ASR this means that dimension-reduction methods are sensitive to differences in training and testing conditions.

In this paper, a fundamentally different principle is proposed for feature *selection* for ASR: to exploit the knowledge implicit in the human auditory system. Importantly, this means the new method does *not* require labeled training data. Humans perform better at speech recognition than machines, particularly for noisy environments. Recently, accurate models of the periphery have become available[7,8]. This motivates the selection of a subset of acoustic features from a larger set by maximizing the similarity of the Euclidian geometry of the selected feature set and the human auditory representation of the signal.

The implementation of our approach relies on perturbation theory. For two features sets to perform similarly in classification, "small" Euclidian distances must be similar in the two domains (except for a scaling). The similarity of "large" distances is immaterial for the classification. The implementation is based on the so-called *sensitivity matrix*, which was first developed in the context of rate-distortion theory[9–11] and has been used for audio coding[12].

The present work is related to the many studies on the usage of auditory models as a front-end for ASR, e.g.,[13–16]. The performance for such front-ends is generally robust to variations in environmental conditions. Importantly, the new approach removes the computational complexity associated with pre-processing the signal with an auditory model. It also avoids the difficulty of formatting the auditory-model output for classification.

A side outcome of our work is that it provides a measure of relative importance of a set of features. In this first study, the most commonly used set of static features, the mel-frequency cepstral coefficients (MFCC)[17], are used. The results confirm that the human auditory model is a good guide for the selection of robust acoustic features. They also show that the initial set of MFCCs corresponds to perceptually important information.

This paper is organized as follows. Sec. 2 discusses a similarity measure for the perceptual and feature domains. Sec. 3 applies the method to ASR. Sec. 4 confirms with experiments that the selected features are effective and robust and Sec. 5, provides conclusions.

2. Maximizing similarity of feature and perceptual domain

Our objective is to select, from a larger set of features, a subset of features that provides a separation of sound classes that is close to that obtained by state-of-the-art auditory models. Ideally, this implies an isometry between the perceptual domain and the selected acoustic feature domain. The mapping from the perceptual domain to the acoustic-feature domain would then be *distance preserving*. To obtain the best approximation to this ideal scenario, we define a new, objective criterion in this section. Thus, we avoid the ad-hoc nature of many auditory-system inspired features.

The motivation for the objective is that human recognition performance indicates that the human auditory periphery provides a relatively good separation of sound classes. We postulate that little information relevant for sound classification is lost in the mapping from the acoustic domain to the *perceptual domain*. It is, however, not clear if the representation is redundant.

2.1. A distance preservation measure

It is not possible to design acoustic features that are a distance-preserving mapping from the perceptual domain. For accurate classification, the preservation of the data geometry near the class boundaries is most critical. The preservation of distances that are short relative to classification boundary curvature is important, whereas the preservation of "long" distances is not important.

Distance measures must be defined in both the perceptual and feature domains. Let $\mathbf{x}_j \in \mathbb{R}^N$ denote the N-dimensional speech signal vector characterizing a segment with time index $j \in \mathbb{Z}$ and let $\hat{\mathbf{x}}_{j,m}$ be a perturbation of $\mathbf{x}_j$ with perturbation index m. A perceptual-domain distortion is defined as a surjective mapping of two signals: $\Upsilon : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+$, where $\mathbb{R}^+$ are the nonnegative reals. Perceptual distortion measures are commonly based on the L^2 norm of the difference between the perceptual-domain signals $\mathbf{y}(\mathbf{x}_j)$ and $\mathbf{y}(\hat{\mathbf{x}}_{j,m})$, where $\mathbf{y} : \mathbb{R}^N \rightarrow \mathbb{R}^K$ is a mapping to the (K-dimensional) perceptual domain, $\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) = \|\mathbf{y}(\mathbf{x}_j) - \mathbf{y}(\hat{\mathbf{x}}_{j,m})\|^2$. This measure is the desired distance measure in the perceptual domain.

A similar distortion measure $\Gamma_i : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+$ can be defined for the feature domain of feature set i. Let $\mathbf{c}_i : \mathbb{R}^N \rightarrow \mathbb{R}^L$ be the mapping from a signal segment $\mathbf{x}_j$ to a set of L features $\mathbf{c}_i(\mathbf{x}_j)$ with set index i. An L^2 norm based measure is then: $\Gamma_i(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) = \|\mathbf{c}_i(\mathbf{x}_j) - \mathbf{c}_i(\hat{\mathbf{x}}_{j,m})\|^2$.

Given a finite sequence of frames $j \in \mathcal{J}$ and a finite set of acoustic perturbations $m \in \mathcal{M}_j$, the distance-preservation objective leads to the objective to find the particular set of features i that minimizes a measure of dissimilarity in the perceptual-domain distortion and the feature-domain distortion,

$$G(i) = \sum_{j \in \mathcal{J}, m \in \mathcal{M}_j} \left[\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) - \lambda \Gamma_i(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})\right]^2, \tag{1}$$

where $\lambda = \frac{\sum_{j\in\mathcal{J},m\in\mathcal{M}_j}\Upsilon(\mathbf{x}_j,\hat{\mathbf{x}}_{j,m})\Gamma_i(\mathbf{x}_j,\hat{\mathbf{x}}_{j,m})}{\sum_{j\in J,m\in M_j}\Gamma_i(\mathbf{x}_j,\hat{\mathbf{x}}_{j,m})^2}$ is an optimal scaling of the acoustic feature criterion. Eq. (1) can be interpreted as a measure of proximity to isometry.

2.2. Perturbation analysis

While it is possible to evaluate Eq. (1) directly even for complex distortion measures, this can be computationally expensive. For short distances, the perceptual distortion measure $\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})$ and the feature-based distortion measure $\Gamma(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})$ can be approximated with simpler quadratic measures, reducing the computational complexity. The approach is based on the sensitivity matrix framework[9–12].

The perturbation analysis for the perceptual domain and the feature domain is identical, and we only describe the first case. For notational brevity we omit the subscripts indicating frame number and perturbation where no ambiguity exists. First, let us consider $\Upsilon(\mathbf{x}, \hat{\mathbf{x}})$ to be known. We assume that $\Upsilon(\mathbf{x}, \mathbf{x}) = 0$ and that this forms a minimum. We furthermore assume that $\Upsilon(\mathbf{x}, \hat{\mathbf{x}})$ is analytic in $\hat{\mathbf{x}}$. Then, for sufficiently small perturbations $\hat{\mathbf{x}} - \mathbf{x}$, we can make the approximation

$$\Upsilon(\mathbf{x}, \hat{\mathbf{x}}) \approx [\hat{\mathbf{x}} - \mathbf{x}]^T \mathbf{D}_\Upsilon(\mathbf{x})[\hat{\mathbf{x}} - \mathbf{x}], \tag{2}$$

where $\mathbf{D}_{\Upsilon,ij}(\mathbf{x}) = \left.\frac{\partial^2\Upsilon(\mathbf{x},\hat{\mathbf{x}})}{\partial\hat{x}_i\partial\hat{x}_j}\right|_{\hat{\mathbf{x}}=\mathbf{x}}$ is the *sensitivity matrix.*

It is common that the mapping from $\mathbf{x}$ to the perceptual or feature domain is given, rather than the distortion criterion. Consider the mapping $\mathbf{c}$ to the feature domain. If the mapping $\mathbf{c}$ is analytic, the Taylor series can be used to make a local approximation around $\mathbf{x}$:

$$\mathbf{c}(\hat{\mathbf{x}}) \approx \mathbf{c}(\mathbf{x}) + \mathbf{A}[\hat{\mathbf{x}} - \mathbf{x}], \tag{3}$$

where $\mathbf{A} = \left.\frac{\partial\mathbf{c}(\mathbf{x})}{\partial\hat{\mathbf{x}}}\right|_{\hat{\mathbf{x}}=\mathbf{x}}$. An L^2 distance measure in the feature domain then leads to a signal domain sensitivity matrix

$$\mathbf{D}_\Gamma(\mathbf{x}) = \mathbf{A}^T\mathbf{A}. \tag{4}$$

Thus, we can write the distortion $\Gamma(\mathbf{x}, \hat{\mathbf{x}})$ in the form of Eq. (2). The sensitivity matrix based expressions facilitate a fast evalution of Eq. (1).

3. Application to speech recognition

In our algorithm the perceptual domain is the domain of the output vectors of the auditory model used. This section illustrates the application of the method to a specific auditory model and specific type of acoustic features.

3.1. van de Par auditory model

The van de Par[8] auditory model is a static psycho-acoustic masking model. As it uses the magnitude spectrum as input, the vector $\mathbf{x}_j$ characterizing speech segment j is now the (square-root) periodogram. The model consists of channels f, in each of which the ratio of the distortion $\hat{\mathbf{x}} - \mathbf{x}$ to masker $\mathbf{x}$ is estimated, where $\mathbf{x}$ denotes the magnitude spectrum of speech. In the end, all ratios are combined together, to account for the spectral integration property of the human auditory system. The complete model is

$$\Upsilon(\mathbf{x}, \hat{\mathbf{x}}) = C_s L_e \sum_{g \in \mathcal{G}} \frac{\frac{1}{N} \sum_{f=0,\cdots,N-1} |h_{om}(f)|^2 |\gamma_i(f)|^2 |x(f) - \hat{x}(f)|^2}{\frac{1}{N} \sum_{f=0,\cdots,N-1} |h_{om}(f)|^2 |\gamma_i(f)|^2 |x(f)|^2 + C_a}, \tag{5}$$

where C_s and C_a are constants calibrated using measurement data, L_e is the effective duration of the segment according to the temporal integration time of the human auditory system, the integer g labels the gamma-tone filter and $\mathcal{G}$ the set of gammatone filters considered, h_{om} is the outer and middle ear transfer function which is the inverse of the threshold in quiet and γ_i is the i'th gammatone filter.

Combining Eq. (2) and Eq. (5), the sensitivity matrix $\mathbf{D}_\Upsilon(\mathbf{x})$ can be obtained. It is a diagonal matrix with the diagonal element for row and column f given by

$$\mathbf{D}_{\Upsilon,ff}(\mathbf{x}) \approx 2 C_s L_e \sum_i \frac{\dfrac{1}{N} \displaystyle\sum_f |h_{om}(f)|^2 |\gamma_i(f)|^2}{\dfrac{1}{N} \displaystyle\sum_f |h_{om}(f)|^2 |\gamma_i(f)|^2 |x(f)|^2 + C_a}. \tag{6}$$

3.2. Local linearization of the MFCCs

In our experiments, the mel-frequency cepstral coefficients (MFCCs)[17] were used since they are the most commonly used acoustic features. The MFCCs are defined as

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \ln\Big\{\sum_{n=0}^{N-1} \mathbf{x}(n)\mathbf{H}_m(n)\Big\}\cos\Big\{q[m-\frac{1}{2}]\frac{\pi}{M}\Big\}, q = 1, \cdots, Q, \tag{7}$$

where $\mathbf{x}(n)$ is the periodogram, $\mathbf{H}_m(n)$ is the m'th triangular mel-filter, m is the filter index, M is the number of triangular bandpass filters used, and Q is the number of cepstrum coefficients.

Sec. 2.2.2 introduced the matrix $\mathbf{A}$ that characterizes the local relation between the features and the signal $\mathbf{x}$. For the MFCCs, the matrix $\mathbf{A}$ is

$$\mathbf{A}_{qn} = \sum_{m=0}^{M-1} \cos\Big\{q[m-\frac{1}{2}]\frac{\pi}{M}\Big\}\frac{\mathbf{H}_m(n)}{\sum_{n=0}^{N-1}\mathbf{x}(n)\mathbf{H}_m(n)}. \tag{8}$$

3.3. Overview of the Algorithm

We now outline the computation of the measure of proximity to isometry. Given an unlabeled database of speech, we compute for each of a large set of frames the periodogram $\mathbf{x}$. We compute the sensitivity matrix $\mathbf{D}_\Upsilon(\mathbf{x})$ using Eq. (6). We also compute the sensitivity $\mathbf{D}_\Gamma(\mathbf{x})$ matrix Eq. (4) using the Jacobian $\mathbf{A}$ of Eq. (8). For each frame we then create a large number of small random perturbations $\hat{\mathbf{x}}$ of $\mathbf{x}$ and evaluate $\Upsilon(\mathbf{x}, \hat{\mathbf{x}})$ using Eq. (2) and $\Gamma(\mathbf{x}, \hat{\mathbf{x}})$ using a corresponding equation. Finally we evaluate Eq. (1) for all perturbations and all frames simultaneously.

4. Experimental results

This section examines the plausability of the linearity assumption used in the perturbation method and verifies the robust performance of the selected feature sets. All experiments were performed on MFCCs.

The MFCCs were extracted from the AURORA2[18] database, sampled at 8 kHz, using

Fig. 1. Scatter plots of the estimated $\delta\mathbf{c}$'s vs. $\delta\mathbf{c}_{true}$'s for the first and second MFCC, respectively.

a Hamming window of 25 ms with an overlap of 12.5 ms. The DFT dimensionality was 256 and the number of filters used was 23. A set of 12 conventional MFCCs was extracted.

4.1. Range of linearity

The range of validity for the linearization assumption between the cepstrum and the speech was examined first. The speech was distorted with i.i.d. Gaussian noise at different SNRs ranging from 30 to 90 dB with a step of one.

Fig. 1 shows the change in the features computed from the linearized relation Eq. (3) versus the true difference between the cepstra of the original and distorted signals, for the first and second MFCC, respectively. The linearity assumption is reasonable at a scale that is meaningful for sound discrimination. The outliers result from regions where the power of the signal is low, as can be seen from Eq. (8).

4.2. Speech recognition experiments

We performed recognition experiments on features derived from the standard set of 12 MFCCs. We compared five types of feature sets for identical dimensionalities $n < 12$. The first set of features results from the auditory-model based feature selection (*amfs*) method introduced in this paper (we use Gaussian perturbations with $|\mathcal{M}_j| = 100$ and an SNR of 100 dB). The second set of features was obtained using standard (homoscedastic) linear discriminant analysis (*lda*)[4]. The third set of features was obtained using standard heteroscedastic linear discriminant analysis (*hlda*)[5]. The average performance of five randomly selected MFCC feature subsets is displayed as *5-rsfs*. The fifth and final set is simply the set of the first n MFCCs, denoted as *initial*.

Note that *lda* and *hlda* have two advantages over *amfs*: *i*) they are *dimension-reduction* methods, rather than *subset-selection* methods and *ii*) they require classified data as training input. The *amfs* method has as advantage that it can rely on knowledge inherent in the auditory periphery.

To build the recognizer we used the HTK[19] toolkit. The digits were modeled as whole word HMMs with 16 states (HTKs notation is 18 states including the beginning and end states) and three Gaussian mixture components per state. To minimize modeling artefacts, the results are for full covariance matrices, but the use of diagonal covariance matrices gives essentially the same results. An initial model with global data means and covariances, identical for each digit, was used and 16 iterations were used to build the final model.

Table 1 shows the recognition accuracy for training and testing on clean data for dimensionality $n = 8$ and for $n = 4$. The caption of the table provides the MFCC subsets selected by *amfs*. For the experiments we added the energy feature "+E". We also performed the experiments with feature sets that were augmented with their velocity ("+V") and acceleration ("+A"). For clarity we note that the subset-selection and dimension-reduction operations were always performed on the static features. Training was performed on the clean training set of 8440 sentences and the testing on the 4004 clean data of test set A.

For clean data the *amfs* selected feature set performs similarly to the *lda* and *hlda* feature sets and to the *initial* set of MFCCS. All these features sets perform significantly better than the average of randomly selected feature sets *5-rsfs*. It is interesting to note that multiple distinct MFCC subsets perform well. Consistent with the recognition results shown in Table 1, the score of the measure of proximity to isometry, given by Eq. (1), is similar for the set *initial* and for the *amfs* selected features.

Table 2 shows the recognition accuracy for noisy data. Again the table caption provides the MFCC subsets selected by *amfs*. The training was performed on the multi-conditioned noisy training set consisting of 6752 files and the testing on the 24024 noisy data of test set A. The results shown in Table 2 are averaged over subway, babble, car, and exhibition additive noise for several SNR values. The *5-rsfs* configuration is the same as for the clean case (the same MFCCs subsets were considered).

For the noisy data, the performance of the *amfs* subsets are in all cases better than *lda*, *hlda* and *5-rsfs*. The performance of *initial* is similar to that of *amfs*, although it does not consistently use the same subset. The results indicate that the new *amfs* method is more robust to environmental noise than *lda* and *hlda*. This increased robustness was confirmed in other experiments where we trained and tested on different environmental conditions. This result is not unexpected as *amfs* is based on an auditory periphery that is robust over a large range of environmental conditions. In contrast, *lda* and *hlda* must rely on the training data only.

Our results indicate that the natural ordering of the MFCCs is perceptually highly relevant. In both Table 1 and Table 2 the initial set of MFCCs, *initial* performs as well as the *amfs* selected set and significantly better than a typical randomly selected set. While *amsfs* does, in general, not simply select the initial set of MFCCs, it always includes the low quefrency MFCCs, indicating that they represent an important component of the perceived information.

5. Conclusions

We conclude that the selection of speech features based on human perception results in robust features that perform well for speech recognition over a range of environmental conditions. Our results suggest that the method results in features that are more robust to noise than either homoscedastic or heteroscedastic discriminant analysis. This implies that effective dimension reduction of feature sets for speech recognition is possible without knowledge of the meaning of the signal (without the availability of classified data). Our results indicate that the human auditory periphery has a parsimonious output representation, as significant redundancy would have made the measure of proximity to isometry, Eq. (1), ineffective for classification.

References and Links

[1] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering”, IEEE Trans. Knowledge Data Eng. **17**, 491502 (2005).

[2] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy”, IEEE Trans. Pat. Analys., Mach. Intellig. **27**, 1226–1238 (2005).

[3] J. H. Plasberg and W. B. Kleijn, “Feature selection under a complexity constraint”, IEEE Trans. Multimedia **11**, 566–571 (2009).

[4] R. Haeb-Umbach and H. Ney, “Linear discriminant analysis for improved large vocabulary continuous speech recognition”, IEEE Int. Conf. Acoust. Speech Sig. Proc. **1**, 13–16 (1992).

[5] N. Kumar and A. Andreou, “Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition”, Speech Communication **267**, 283–297 (1998).

[6] F. Valente and C. Wellekens, “Maximum entropy discrimination (MED) feature subset selection for speech recognition”, IEEE Works. on ASRU 327–332 (2003).

[7] T. Dau, D. Püschel, and A. Kohlrausch, “A quantitative model of the ”effective” signal processing in the auditory system. I. Model structure”, J. Acoust. Soc. Am. **99**, 3615 – 3622 (1996).

[8] S. van de Par, G. Charestan, and R. Heusdens, “A gammatone-based psychoacoustical modeling approach for speech and audio coding”, Proc.ProRISC,Veldhoven,NL 321–326 (2001).

[9] W. R. Gardner and B. D. Rao, “Theoretical analysis of the high-rate vector quantization of LPC parameters”, IEEE Trans. Speech, Audio Proc. **3**, 367–381 (1995).

[10] T. Linder, R. Zamir, and K. Zeger, “High-resolution source coding for non-difference distortion measures: multidimensional companding”, IEEE Trans. Inform. Theory **45**, 548–561 (1999).

[11] J. Li, N. Chaddha, and R. M. Gray, “Asymptotic performance of vector quantizers with

a perceptual distortion measure", IEEE Trans. Inform. Theory **45**, 1082–1091 (1999).

[12] J. H. Plasberg and W. B. Kleijn, "The sensitivity matrix: Using advanced auditory models in speech and audio processing", IEEE Transactions on Audio, Speech and Language Processing **15**, 310–319 (2007).

[13] S. Seneff, "A joint synchrony/mean rate model of auditory speech processing", J. Phonet. **16**, 55–76 (1988).

[14] O. Ghitza, "Auditory nerve representation as a basis for speech processing", in *Advances in Speech Signal Process.*, 453–485 (Marcel Dekker) (1991).

[15] W. Jeon and B.-H. Juang, "A study of auditory modeling and processing for speech signals", volume 1, 929 – 932 (2005).

[16] S. Haque, R. Togneri, and A. Zaknich, "A temporal auditory model with adaptation for automatic speech recognition", volume 4, 1141–1144 (2007).

[17] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoust. Speech Sig. Proc. **28**, 357–366 (1980).

[18] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions", in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millenium* (Paris) (2000).

[19] S. Young and et al, *The HTK Book (for HTK Version 3.2)* (Cambridge University, Engineering Department) (2002).

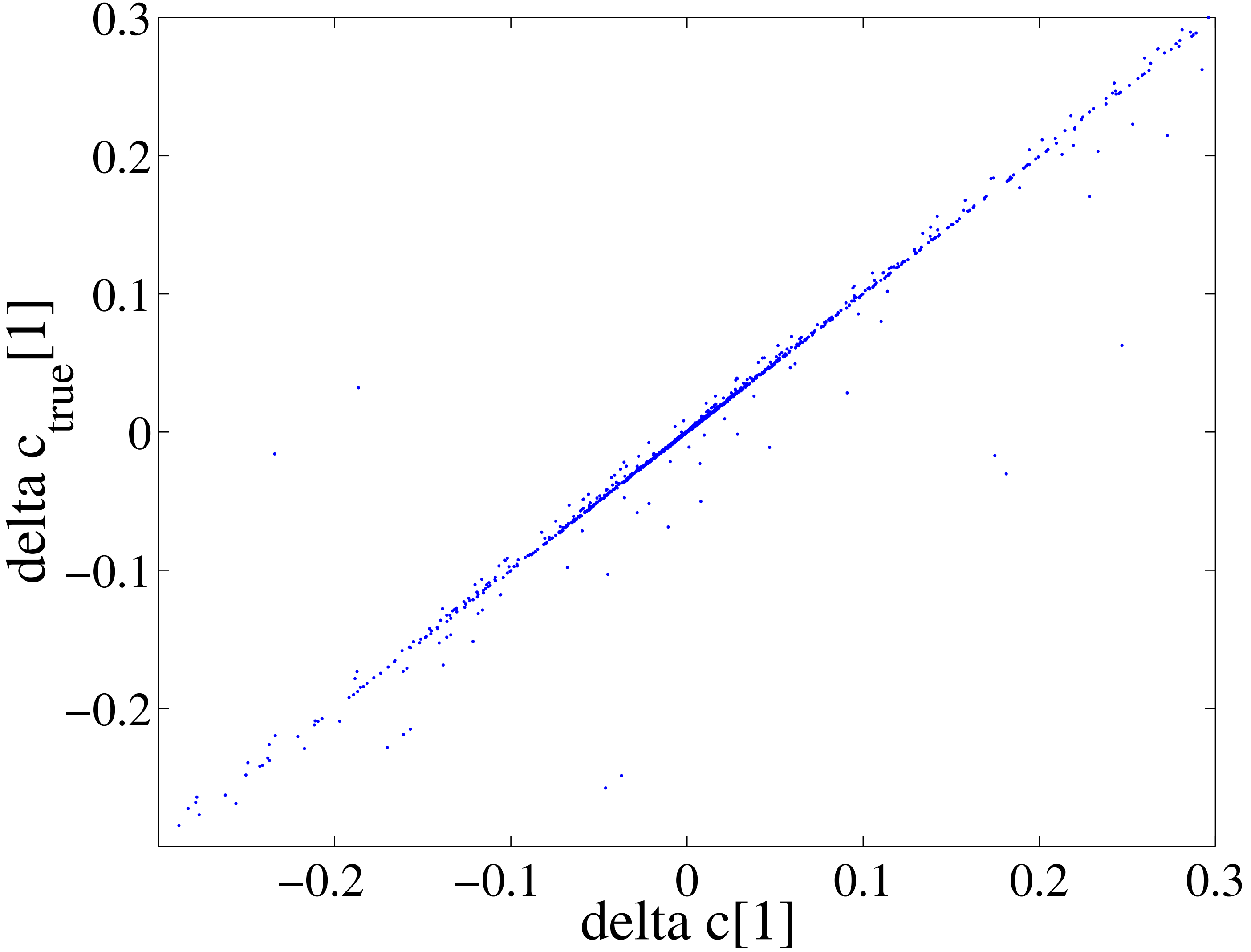
Table 1. AURORA2 clean training results. The *amfs* selected coefficients with indices *1,2,3,4,9,10,11,12* for n=8 and *1,2,3,12* for n=4, respectively.

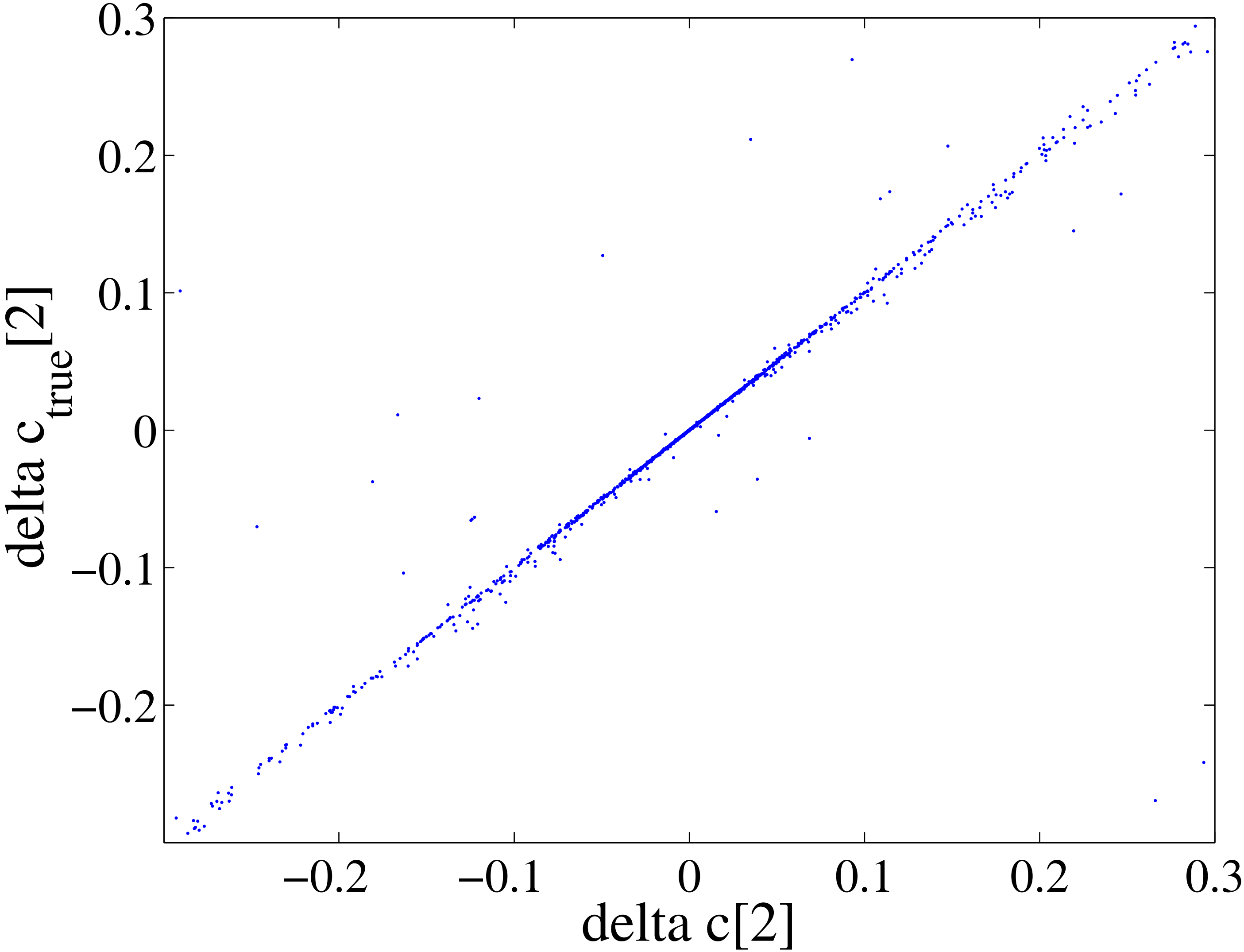
		Data Test Set A				
feature set		*clean 1*	*clean 2*	*clean 3*	*clean 4*	*avg.*
12+E	*full,static*	98.1	97.6	97.7	98.1	**97.9**
12+E+V+A	*full,dynamic*	98.9	99.1	98.8	99.2	**99.0**
8+E	*amfs*	97.8	97.6	97.1	97.8	**97.6**
	lda	98.3	97.8	98.4	98.3	**98.2**
	hlda	98.7	98.1	98.5	99.0	**98.6**
	5-rsfs	95.8	95.7	95.5	96.0	**95.8**
	initial	97.5	97.3	97.6	97.9	**97.6**
8+E+V+A	*amfs*	99.1	98.9	98.8	99.2	**99.0**
	lda	98.4	98.1	98.2	98.6	**98.3**
	hlda	98.6	98.5	98.5	99.1	**98.7**
	initial	98.9	98.9	98.8	99.1	**98.9**
4+E	*amfs*	96.9	96.5	96.6	97.3	**96.8**
	lda	96.3	95.5	95.7	96.7	**96.1**
	hlda	95.9	95.0	96.0	96.5	**95.9**
	5-rsfs	85.5	85.3	85.6	85.6	**85.5**
	initial	97.0	96.5	96.7	97.3	**96.9**
4+E+V+A	*amfs*	98.6	98.9	98.7	99.0	**98.8**
	lda	98.3	97.9	98.2	98.2	**98.2**
	hlda	98.7	98.3	98.3	98.9	**98.6**
	initial	98.6	98.8	98.6	99.2	**98.8**

Table 2. AURORA2 multi-conditioning noisy training results. The *amfs* selected coefficients with indices *1,2,3,4,5,6,11,12* for n=8 and *1,2,3,4* for n=4, respectively.

		Data Test Set A				
feature set		*20 dB*	*15 dB*	*10 dB*	*5 dB*	*0 dB*
12+E	*full,static*	94.8	93.1	88.4	75.0	47.4
12+E+V+A	*full,dynamic*	97.4	96.5	94.0	86.0	59.6
8+E	*amfs*	93.6	91.4	85.7	65.1	34.9
	lda	84.0	78.1	63.2	40.8	17.8
	hlda	85.0	80.9	71.7	55.7	33.2
	5-rsfs	88.0	84.6	75.6	53.0	24.6
	initial	93.1	91.0	86.0	69.2	38.1
8+E+V+A	*amfs*	97.1	96.0	93.4	84.1	56.8
	lda	94.4	92.8	87.8	76.5	52.0
	hlda	94.8	93.4	89.4	79.3	58.4
	initial	97.5	96.1	92.9	83.1	53.2
4+E	*amfs*	91.1	87.4	76.9	50.9	20.9
	lda	40.5	32.5	19.6	9.9	7.7
	hlda	44.0	38.8	28.6	16.6	8.4
	5-rsfs	66.3	59.3	43.6	24.0	11.9
	initial	91.1	87.4	76.9	50.9	20.9
4+E+V+A	*amfs*	96.0	94.6	90.9	79.4	50.1
	lda	93.6	92.1	86.3	73.9	48.9
	hlda	94.2	92.8	88.9	78.8	58.5
	initial	96.0	94.6	90.9	79.4	50.1

List of Figures


0.3
0.2
0.1
0
−0.1
−0.2
delta c$_{true}$[1]
−0.2
−0.1
0
0.1
0.2
0.3
delta c[1]


0.3
0.2
0.1
0
−0.1
−0.2
$\text{delta } c_{\text{true}}[2]$
−0.2
−0.1
0
0.1
0.2
0.3
delta c[2]