2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust

Dimensionality Reduction for Emotional Speech Recognition

Pouria Fewzee, Fakhri Karray
Centre for Pattern Analysis and Machine Intelligence
University of Waterloo, Waterloo, ON, Canada N2L 3G1
Email: {fewzee,karray}@pami.uwaterloo.ca

Abstract

The number of speech features that are introduced to emotional speech recognition exceeds some thousands and this makes dimensionality reduction an inevitable part of an emotional speech recognition system. The elastic net, the greedy feature selection, and the supervised principal component analysis are three recently developed dimensionality reduction algorithms that we have considered their application to tackle this issue. Together with PCA, these four methods include both supervised and unsupervised, as well as filter and projection-type dimensionality reduction methods. For experimental reasons, we have chosen VAM corpus. We have extracted two sets of features and have investigated the efficiency of the application of the four dimensionality reduction methods to the combination of the two sets, besides each of the two. The experimental results of this study show that in spite of a dimensionality reduction stage, a longer vector of speech features does not necessarily result in a more accurate prediction of emotion.

1. Introduction

Emotional speech recognition is aimed at enabling computers to predict the emotional state of speakers. To pursue this objective, one may start with identifying a set of useful speech features; useful in the sense that the set of features can effectively extract the emotional content of speech. When we say *effectively*, we intend to bring into attention two qualities of a set of features: (1) to lead to a high prediction accuracy, and (2) to be minimal in size. These two happen to be the objectives of dimensionality reduction. In this work, we focus on the dimensionality reduction aspects of affective speech recognition.

The number of speech features that are being put into practice for emotional speech recognition is increasing. Recently, a vector of about 2K dimensions has been set as the standard for the first international audio/visual challenge [1]. Moreover, the tendency towards the investigation of new speech features [2], [3], [4], [5] seems to be ongoing. Now, the question is how hard of a problem it is to select d out of 2000 features, when the optimal value for d is unknown. For this many features, there are 2^{2000} possible combinations, which is of an order of 10^{600}. Computationally speaking, evaluating this many combinations of features is out of question.

The problem of dimensionality reduction has been frequently addressed in the literature of emotional speech recognition. For instance, in the category of filter methods, Schuller [6] has used correlation based feature selection (CFS). From the same category, Kim and others [2] have adopted information theoretic subset selection methods. From the category of wrappers, Sun and Moore [5] have made use of sequential forward selection (SFS). Also, the application of principal component analysis (PCA) has been frequently seen as a common extraction-type dimensionality reduction method [3], [7].

In this work, we investigate the application of three fairly recent dimensionality reduction methods to the affective speech recognition. Those are supervised principal component analysis (SPCA) [8], greedy feature selection (GFS) [9], and the elastic net [10]. Therefore, to see where they stand compared to other dimensionality reduction algorithms, we have included the application of PCA in our work. This subset of dimensionality reduction methods cover both unsupervised (i.e. PCA and the GFS) and supervised methods (i.e. SPCA and the elastic net). Also, it includes both filter (i.e. GFS and the elastic net) and extraction (i.e. PCA and SPCA) methods.

The remainder of this work is organized as follows. In Section 2, we review different approaches to dimensionality reduction. To do so, we talk about those aspects that they share, as well as those that make them different. Then, we present the results of this

978-0-7695-4848-7/12 $26.00 © 2012 IEEE
DOI 10.1109/SocialCom-PASSAT.2012.83

study in Section 3. This work is brought to an end by concluding remarks in Section 4.

2. Dimensionality Reduction

Given the explanatory variables $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\} = \mathbf{X} \in \mathcal{X}$ and the response variable $\mathbf{y}$, the objective of dimensionality reduction is to find a subspace $\hat{\mathcal{X}} \subseteq \mathcal{X}$ with minimal dimensionality d, which can satisfy a particular criterion. In the case of our problem, which is supervised learning, the criterion is to maximize the prediction accuracy.

Regardless of the nature of a learning problem, which can be either supervised or unsupervised, in the search for the subspace $\hat{\mathcal{X}}$, either to take the response variable $\mathbf{y}$ into consideration or not, is the matter of supervised or unsupervised dimensionality reduction methods. From another point of view, there are two major approaches to the problem: variable selection and projection (a.k.a. feature extraction). While the former reduces the covariates' space to a subset of existing covariates, the latter gives out a combination of those as a solution. These two categories can be unified by the following equation.

$$\hat{\mathbf{X}} = \mathbf{X}\mathbf{W} \tag{1}$$

Where for selection methods, each of the columns of $\mathbf{W}$ have all zero entries except one; the non-zero entry indicates the index of a selected variable. Whereas for projection methods, entries of $\mathbf{W}$ may take values from all over the range.

Depending on one's perspective, one of these two approaches might be preferable over the other. On the one hand, the selection algorithms preserve the nature of the original variables, and that makes the interpretation of a model more feasible. On the other hand, sparsity of the transformation matrix $\mathbf{W}$, which is the case for selection methods, may not be easily achievable.

In this work, we investigate the application of four different dimensionality reduction algorithms to emotional speech recognition: supervised principal component analysis (SPCA) [8], greedy feature selection (GFS) [9], elastic net [10], and principal component analysis (PCA). This subset of dimensionality reduction methods is chosen due to the variety of approaches that it comprises. SPCA and elastic net are supervised methods, whereas greedy feature selection and PCA are unsupervised methods. On the other hand, greedy feature selection and elastic net can be used for selection, whereas SPCA and PCA are projectors. Furthermore, the application of SPCA and greedy feature selection has not yet been investigated in the speech community. Also, elastic net [11] has not yet been comprehensively explored in this community.

3. Experiments

To investigate the application of the four dimensionality reduction algorithms we perform some experiments. In this section, we talk about our choice of corpus and speech features. To model emotional content of speech with respect to the extracted features, we use maximum likelihood estimator, having assumed that the prediction error is Gaussian. We then represent and discuss the results of the experiments. Finally, we show where the results of this study stand, compared to a recent work on the same corpus.

3.1. Database

Our choice of database in this work is VAM. VAM [12] is a spontaneous emotional speech database. The database is composed of 12 hours of recording, available in both audio and visual signals. A total number of 104 speakers take part in the recordings.The database is annotated using three emotional primitives: valence, activation, and dominance. VAM is split into two parts, VAM I and VAM II, and we use the two parts together (VAM I+II).

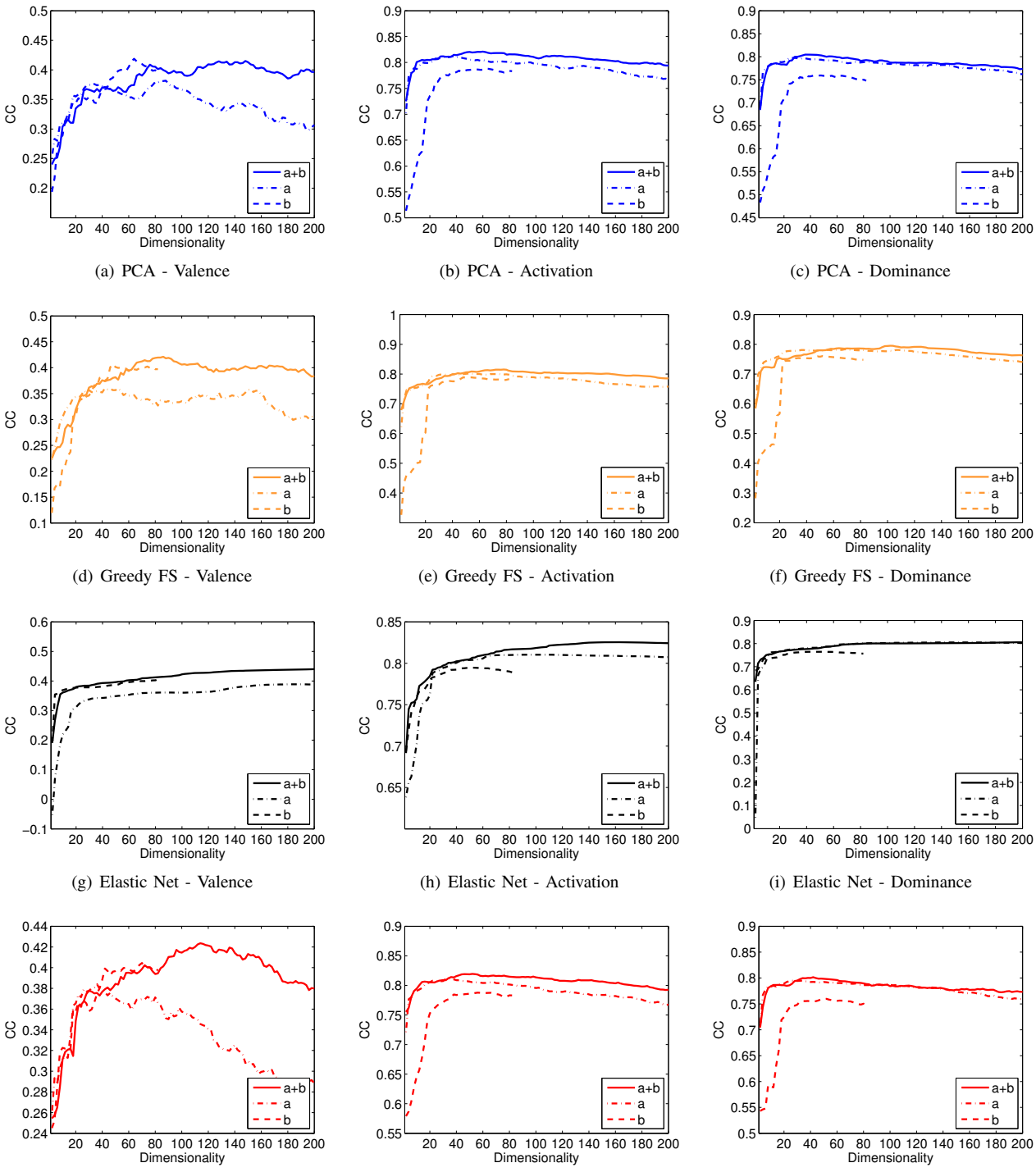
3.2. Speech Features

We have extracted two sets of features in this work.

set `a`. This is composed of a set of features that we call spectral energy distribution (SED). SED is made of a number of components. For a speech signal $s[n]$, the component i is defined as follows.

$$SED_s^i = \sum_{l_i}^{u_i} g(|S(K)|^2). \tag{2}$$

Where $S[k]$ denotes the discrete Fourier transform of the signal $s[n]$ and $[l_i, u_i]$ specifies the corresponding spectral interval to the component i. In this work, we have set the parameters so that $l_1 = 0Hz$, $l_i = u_{i-1}$, and $u_i - l_i = 100Hz$. The components cover the whole spectral range (0-8KHz according to the database of interest). The function $g(.)$ (Eq. 2) is set to the family of rational exponents, with exponents of 0.15, 0.2, and 0.3 for each one of the valence, activation, and dominance dimensions. SED is extracted from 100 msec windows, and as for the statistics, we have used the minimum, maximum, mean, median, and



(a) PCA - Valence (b) PCA - Activation (c) PCA - Dominance

(d) Greedy FS - Valence (e) Greedy FS - Activation (f) Greedy FS - Dominance

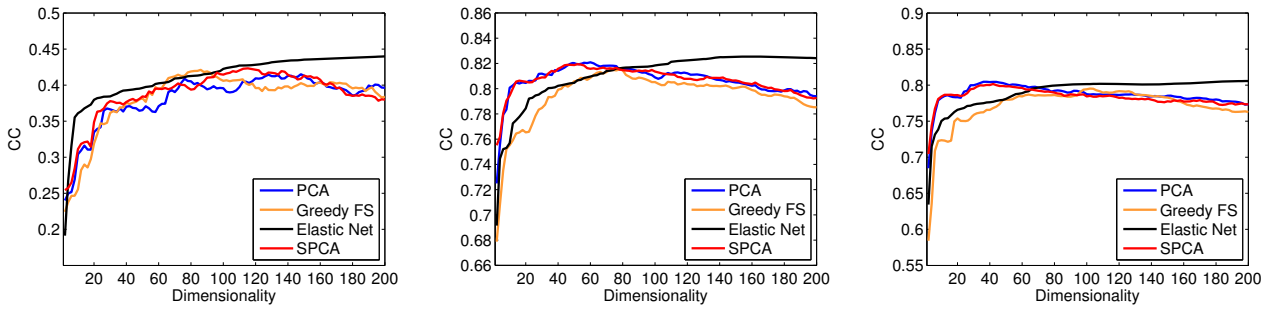
(g) Elastic Net - Valence (h) Elastic Net - Activation (i) Elastic Net - Dominance

(j) SPCA - Valence (k) SPCA - Activation (l) SPCA - Dominance

Figure 1. Dimensionality reduction results (CC: correlation coefficient)

Feature Set	Reduction Method	Valence CC(MAE)	NF	Activation CC(MAE)	NF	Dominance CC(MAE)	NF	**mean** CC(MAE)	NF
a+b	PCA	0.42(0.14)	148	0.82(0.16)	**60**	0.80(0.14)	**38**	0.68(0.15)	82
	Greedy FS	0.42(0.14)	**86**	0.82(0.16)	78	0.80(0.14)	102	0.68(0.15)	89
	Elastic Net	**0.44(0.13)**	191	**0.83(0.15)**	156	**0.81(0.14)**	199	**0.69(0.14)**	182
	SPCA	0.42(0.14)	114	0.82(0.15)	**52**	0.80(0.14)	**42**	0.68(0.14)	**69**
a	PCA	0.38(0.14)	88	0.81(0.16)	**38**	0.80(0.14)	**26**	0.66(0.15)	51
	Greedy FS	0.37(0.14)	**42**	0.80(0.16)	52	0.78(0.17)	70	0.65(0.16)	55
	Elastic Net	**0.39(0.14)**	167	0.81(0.16)	92	**0.81(0.14)**	145	**0.67(0.15)**	135
	SPCA	0.38(0.14)	**36**	0.81(0.16)	**34**	0.80(0.14)	34	0.66(0.15)	**34**
b	PCA	**0.42(0.14)**	64	0.79(0.17)	62	0.76(0.15)	**36**	**0.66(0.15)**	54
	Greedy FS	0.41(0.14)	**48**	0.79(0.17)	**48**	0.76(0.15)	52	0.65(0.15)	**49**
	Elastic Net	0.41(0.14)	78	0.79(0.17)	**49**	0.76(0.15)	45	0.65(0.15)	57
	SPCA	0.40(0.14)	70	0.79(0.17)	56	0.76(0.15)	52	0.65(0.15)	59

Table 1. Summary of dimensionality reduction results (CC: correlation coefficient, MLE: mean absolute error, NF: number of features)



(a) Valence (b) Activation (c) Dominance

(d) Valence (e) Activation (f) Dominance

Figure 2. Dimensionality reduction results for the a+b feature set (CC: correlation coefficient, MLE: mean absolute error)

the variance. The total number of features in this set is 400.

set b. The features in this set are the fundamental frequency, the first three formants, the first twelve MFCCs, total energy, and the zero cross-over rate. For the fundamental frequency, formants, and MFCCs extraction is done from 50 mSec windows of signal; we have then computed the minimum, maximum, mean, median, and variance. The total number of features in this set adds up to 82.

For experimental purposes, we use each of the two feature sets a and b, as well as their combination, which we denote by a+b.

3.3. Results

10-fold cross validation (CV) is set as a standard [6], [12] for evaluating prediction result on the VAM database and we adopt to that. For the sake of fairness of the comparisons that we are going to make

among the four algorithms, we will fix the CV indices throughout the experiment. Pearson's correlation coefficient (CC) and mean absolute error (MAE, also referred to as mean linear error or MLE) has been used as the means of evaluation of the prediction accuracy.

The result of experiments is shown on Figures 1 and 2, and Table 1.

3.4. Discussion

According to Figure 1, we can see that, regardless of the choice of dimensionality reduction method, the feature set a+b, although contains all the features from feature sets a and b, does not necessarily lead to a better prediction accuracy, compared to the scenarios that we used just one of the two. In other words, from an optimization point of view, all the four dimensionality reduction methods that we used in this study are likely to suffer from sub-optimality. In general, we can see that for the relatively smaller number of features, usually the feature sets a and b result in a better prediction accuracy than that of the a+b. Now, let us take a look at a few different cases of sub-optimality of solutions, based on the results presented on Figure 1.

- According to the Figure 1(a), the best accuracy obtained resulting from dimensionality reduction by PCA belongs to the feature set b. This means that regardless of the dimensionality, PCA could not find a transformation of the a+b space that can result in a prediction as accurate as that of the b space.
- For the valence dimension and dimensionalities less than 60, both greedy feature selection and SPCA's most accurate prediction is obtained by one of the feature sets a or b.
- When we used the elastic net for the dominance dimension (Figure 1(i)), the accuracy of the prediction using the feature set a, although not significantly, outperforms that of the feature set a+b, for the most part of the dimensionality range.

Now, we would like to compare the prediction accuracy resulting from dimensionality reduction by the four algorithms, for each of the emotional primitives (Table 1). For the valence dimension, for feature sets a and a+b, elastic net's reduction leads to the best accuracy, although it takes the highest number of dimensions among all to obtain that accuracy. For the same feature sets, the accuracy of the prediction resulting from reduction by the greedy feature selection is comparable to that of the elastic net, but it takes the least dimensionality for the greedy feature selection to do the job. For feature set b, however, PCA's reduction gives the most accurate prediction among all the four algorithms. For the activation dimension, for the feature set a+b, elastic net's reduction results in the most accurate prediction, which again comes at the price of the highest number of dimensions. For the same dimension and feature set, SPCA results in a comparable accuracy to that of the elastic net, however with way less number of dimensions (33%). For the dominance emotion primitive, for the feature set a+b, elastic net's reduction results in the most accurate prediction, and again with the highest number of dimensions. For the same dimension and feature set, both PCA and SPCA result in comparable prediction accuracies to that of the elastic net, however with significantly less number of dimensions (about 20%).

3.5. A Comparison

To see where the results of this work stand compared to recent works on the same database, Table 2 puts two sets of results of this work side-by-side with those of a work by Schuller [6]. The first set is the one which has the most accurate predictions (elastic net, feature set a+b) and the second set has the next most accurate predictions and at the same time the least number of features (SPCA, feature set a+b). Based on the results presented on this table, we can see that the average prediction accuracy of the elastic net is higher than the other two, however the dimensionality of the feature vector used for this task is considerably greater than those. On the other hand, SPCA offers a relatively accurate prediction accuracy, but the shortest feature vector, compared to the other two.

4. Conclusion

In this work, we investigated the application of three recent dimensionality reduction algorithms, namely the greedy feature selection, the elastic net, and SPCA, to the emotional speech recognition problem. We extracted two sets of speech features from the VAM corpus and conducted experiments based on each of the two sets, as well as their combination. According to the results of this study, we conclude that, despite of taking dimensionality reduction into consideration, due to the suboptimality of the dimensionality reduction algorithms, a longer vector of features does not necessarily lead to a more accurate prediction. Also, the quality of the choice of dimensionality reduction algorithm, in terms of the resulting accuracy and the size of the target space, depends on the specifications of the learning task. In this work we studied three different

	Valence		Activation		Dominance		**mean**	
	CC(MAE)	NF	CC(MAE)	NF	CC(MAE)	NF	CC(MAE)	NF
Schuller [6]	**0.45(0.13)**	238	0.81(0.16)	109	0.79(0.14)	88	0.68(0.14)	145
this work (elastic net)	0.44(0.13)	191	**0.83(0.15)**	156	**0.81(0.14)**	199	**0.69(0.14)**	182
this work (SPCA)	0.42(0.14)	**114**	0.82(0.15)	**52**	0.80(0.14)	**42**	0.68(0.14)	**69**

Table 2. A comparison (CC: correlation coefficient, MLE: mean absolute error, NF: number of features)

learning tasks (valence, activation, and dominance), using three different feature sets (`a`, `b`, and `a+b`), and we can not confidently say that one approach is superior to the others, from both accuracy and efficiency points of view.

Acknowledgment

The authors would like to thank Ahmed Farahat and Elnaz Barshan for providing them with their programs for greedy feature selection and supervised principal component analysis, respectively.

References

[1] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 415–424.

[2] J. Kim, H. Rao, and M. Clements, "Investigating the use of formant based features for detection of affective dimensions in speech," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 369–377.

[3] H. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on hidden markov models," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 378–387.

[4] A. Sayedelahl, P. Fewzee, M. Kamel, and F. Karray, "Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 407–414.

[5] R. Sun and E. Moore, "Investigating glottal parameters and teager energy operators in emotion recognition," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 425–434.

[6] B. Schuller, "Recognizing affect from linguistic information in 3d continuous space," *Affective Computing, IEEE Transactions on*, vol. 2, no. 4, pp. 192 –205, oct.-dec. 2011.

[7] S. Pan, J. Tao, and Y. Li, "The casia audio emotion recognition method for audio/visual emotion challenge 2011," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 388–395.

[8] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recognition*, vol. 44, no. 7, pp. 1357 – 1371, 2011.

[9] A. Farahat, A. Ghodsi, and M. Kamel, "An efficient greedy method for unsupervised feature selection," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, dec. 2011, pp. 161 –170.

[10] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[11] Q. F. Tan, P. Georgiou, and S. Narayanan, "Enhanced sparse imputation techniques for a robust speech recognition front-end," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2418–2429, Nov. 2011.

[12] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-visual emotion recognition using an emotion space concept," *Signal Processing*, 2008.