MULTI-VIEW CCA-BASED ACOUSTIC FEATURES FOR PHONETIC RECOGNITION ACROSS SPEAKERS AND DOMAINS

Raman Arora and Karen Livescu

Toyota Technological Institute at Chicago (TTIC), Chicago, IL 60637

arora@ttic.edu, klivescu@ttic.edu

ABSTRACT

Canonical correlation analysis (CCA) and kernel CCA can be used for unsupervised learning of acoustic features when a second view (e.g., articulatory measurements) is available for some training data, and such projections have been used to improve phonetic frame classification. Here we study the behavior of CCA-based acoustic features on the task of phonetic recognition, and investigate to what extent they are speaker-independent or domain-independent. The acoustic features are learned using data drawn from the University of Wisconsin X-ray Microbeam Database (XRMB). The features are evaluated within and across speakers on XRMB data, as well as on out-of-domain TIMIT and MOCHA-TIMIT data. Experimental results show consistent improvement with the learned acoustic features over baseline MFCCs and PCA projections. In both speaker-dependent and cross-speaker experiments, phonetic error rates are improved by 4-9% absolute (10-23% relative) using CCA-based features over baseline MFCCs. In cross-domain phonetic recognition (training on XRMB and testing on MOCHA or TIMIT), the learned projections provide smaller improvements.

***Index Terms*—** multi-view learning, canonical correlation analysis, articulatory measurements, XRMB, MOCHA-TIMIT, TIMIT, speaker-independence, domain-independence

1. INTRODUCTION

A common approach to acoustic feature vector generation for speech processing tasks is to first construct a high-dimensional acoustic feature vector by concatenating multiple consecutive frames of raw features (e.g., MFCCs or PLPs), and then to reduce dimensionality using a feature transformation. The transformation may be an unsupervised one such as principal components analysis (PCA), a linear supervised transformation such as linear discriminant analysis (LDA) and its extensions [1, 2], or a nonlinear supervised transformation [3]. In this work we consider unsupervised transformation learning, but in a setting where a second "view" of the speech data is available for some training data. In particular, we consider the case where articulatory measurements are available as training data, but not at test time, and ask whether we can use the articulatory information to learn which directions in the acoustic space are most useful. The approach we present avoids some of the disadvantages of unsupervised approaches, such as PCA, which are sensitive to noise and data scaling, and possibly of supervised approaches, which are task-specific.

Articulatory information has been used in speech recognition in a number of ways [4]. Several databases of simultaneous acoustic and articulatory recordings are available (e.g., [5, 6, 7]). Phonetic classification and recognition can be improved if such articulatory measurements are available at test time [8, 9]. This is an unnatural setting, but it suggests that in the absence of articulatory data at test time, perhaps we can predict the articulation from acoustics and use the predicted values as additional observations. However, acoustic-to-articulatory inversion is a complex task (e.g., [10, 11]), and to date it has been difficult to improve recognition performance in this way [8, 9]. Other approaches use articulatory data at training time and attempt to leave it hidden (i.e., implicitly predict it) at test time [12]. Alternatively, knowledge-based approaches, in which articulatory information is never measured but rather used to constrain the hidden state structure, have also been proposed [13, 14, 15, 16].

While predicting articulation may be difficult, learning acoustic features that are somehow *informed* by articulation may be easier. In previous work [17, 18], we have shown that this is indeed possible, and can be used to improve phonetic frame classification. This work applies ideas from *multi-view learning*, in which multiple views of the data are available for training but possibly not for testing [19].

Our approach is based on canonical correlation analysis (CCA), which finds pairs of maximally correlated linear projections of data in two views [20], and its nonlinear counterpart kernel CCA [21]. The two views are the acoustic and articulatory data, and only the acoustic projections are used at test time. The intuition is that articulatory measurements provide information about the linguistic content, and that much of the non-discriminative information in the two views is

The authors thank Soham De and Victoria Evelkin for help with data preparation. This research was supported by NSF grant IIS-0905633. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

largely uncorrelated and therefore filtered out. CCA/KCCA have also been used with audio and video for speaker clustering [22] and identification [23]; for speaker normalization [24], where the views are the speakers; for articulatory inversion [25]; and to study critical articulators [26].

In this paper our goal is to study the applicability of the multi-view acoustic feature learning approach to more practical tasks and settings. We begin by applying the approach to phonetic recognition. Here there are two training phases – a feature learning phase, where both views are used, and a recognizer training phase, where only the acoustics are used. To our knowledge, this is the first time CCA- or KCCA-based features have been used for this task. In addition, in order to be practical, the approach should be applicable to new speakers and domains, for which no articulatory training data exists. We study the degree to which the learned features are speaker-independent or domain-independent, including independence of dialect, recording conditions, and so on.

2. METHODS

The methods used here are based on those of [18]. We briefly review them for completeness. CCA and KCCA are techniques for learning linear or nonlinear functions of data in two views that are maximally correlated [20, 21]. Linear discriminant analysis (LDA) is a special case of CCA where one of the views is the labels. Unlike PCA, CCA is scale-invariant. One assumption typically made is that the two views are largely uncorrelated conditioned on some class of interest (in our case the phonetic class), so that the dimensions that are correlated between the two views should be discriminative for classification. This assumption is imperfect, and some work has recently been devoted to relaxing it [27].

2.1. CCA and kernel CCA

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the spaces of vectors in the two views, $\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}$ the Reproducing Kernel Hilbert Spaces (RKHS) of functions on $\mathcal{X}, \mathcal{Y}$, and k_x, k_y the associated positive definite kernels. We consider random vectors $X \in \mathcal{X}, Y \in \mathcal{Y}$, with an unknown joint distribution that we can access only through N training instances, $\{x_i, y_i\}_{i=1}^N$. In our case, each pair (x_i, y_i) represents features computed over one frame of simultaneously recorded acoustics (x_i) and articulation (y_i).

Kernel CCA finds pairs of nonlinear projections of the two views. The first pair of projections is defined as those functions $f_1 \in \mathcal{H}_x, g_1 \in \mathcal{H}_y$ that solve the optimization problem

$$\{f_1, g_1\} = \operatorname*{arg\,max}_{f \in \mathcal{H}_x, g \in \mathcal{H}_y} \frac{\mathrm{cov}\,(f(X), g(Y))}{\sqrt{\mathrm{var}\,(f(X)) \cdot \mathrm{var}\,(g(Y))}}, \qquad (1)$$

i.e., that maximize correlation between $f(X)$ and $g(Y)$. Subsequent projections $\{f_j, g_j\}$ for $j > 1$ are found by solving (1) subject to the constraints that $f_j(X)$ is uncorrelated with $f_i(X)$, $g_j(Y)$ is uncorrelated with $g_i(Y)$ and $f_j(X)$ is uncorrelated with $g_i(Y)$ for all $i \neq j$. CCA solves the same problem for the case where the projections are linear, $f(X) = v^T X, g(Y) = w^T Y$. CCA and KCCA can be used for dimensionality reduction, by keeping the top projections.

For CCA, the solution is straightforward [21]: The vectors v that maximize the objective are the top eigenvectors of $C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx}$, and w are given as $w \propto C_{yy}^{-1} C_{yx} v$, where C_{xx}, C_{yy} are the autocovariance matrices in each view and C_{xy} is the cross-covariance matrix between X and Y.

To solve the nonlinear KCCA problem, we use the "kernel trick" [28]. Since the nonlinear maps $f \in \mathcal{H}_x$, $g \in \mathcal{H}_y$ are in RKHS, we can express them as linear combinations of the kernel evaluated at the data: $f(x) = \sum_{i=1}^N \alpha_i k_x(x, x_i)$, and similarly for $g(y)$. KCCA can then be written as finding directions $\alpha_1, \beta_1 \in \mathbb{R}^N$ that solve the optimization problem

$$\{\alpha_1, \beta_1\} = \operatorname*{arg\,max}_{\alpha \in \mathbb{R}^N, \beta \in \mathbb{R}^N} \frac{\alpha^T K_x K_y \beta}{\sqrt{(\alpha^T K_x^2 \alpha)\ (\beta^T K_y^2 \beta)}}, \qquad (2)$$

where $K_x \in \mathbb{R}^{N \times N}$ is the centered Gram matrix $K_x = K - K\mathbf{1} - \mathbf{1}K + \mathbf{1}K\mathbf{1}$, $K_{ij} = k_x(x_i, x_j)$ and $\mathbf{1} \in \mathbb{R}^{N \times N}$ is an all-1s matrix, and similarly for K_y. Subsequent vectors $\{\alpha_j, \beta_j\}$ are solutions of (2) with the constraints that the resulting $\{f_j(X), g_j(Y)\}$ are uncorrelated with the previous ones. To alleviate over-fitting, one instead typically maximizes the regularized objective [21]

$$\frac{\alpha^T K_x K_y \beta}{\sqrt{(\alpha^T K_x^2 \alpha + r_x \alpha^T K_x \alpha)\ (\beta^T K_y^2 \beta + r_y \beta^T K_y \beta)}}. \qquad (3)$$

where r_x, r_y are regularization parameters. The optimization is in principle simple: The objective is maximized by the top eigenvectors of the matrix

$$(K_x + r_x I)^{-1}\ K_y\ (K_y + r_y I)^{-1}\ K_x, \qquad (4)$$

In practice this is not straighforward, since the kernel matrices may be too large to compute the eigenvectors or even to construct the matrix in Eq. 4. In previous work [18] we have addressed this issue, and we next summarize the approach.

2.2. Scalable KCCA

As has been observed by ourselves and others [22], it can be useful to further regularize by first projecting the data onto an intermediate-dimensionality space, between the target and original dimensionality. This is especially true for KCCA, where the matrices involved grow with the number of training examples. For KCCA, the intermediate-dimensionality projection can be done by decomposing the kernel matrix as a gram-product of two lower-dimensional matrices:

$$K_x \approx F^T F, \ \ K_y \approx G^T G, \qquad (5)$$

where $F, G \in \mathbb{R}^{m \times N}$, for an intermediate dimensionality $m \ll N$. These form lower-dimensional representations of the maps f, g. Next, let

$$\begin{aligned} C_{ff} &= FF^T, \quad C_{gg} = GG^T, \\ C_{fg} &= FG^T, \quad C_{gf} = GF^T. \end{aligned} \tag{6}$$

The KCCA directions $(\hat{\alpha}, \hat{\beta})$ in the reduced dimensionality are related to the true KCCA directions (α, β) via $\hat{\alpha} = F\alpha$ and $\hat{\beta} = G\beta$. As in CCA, the reduced dimensionality KCCA directions are solutions to the eigenvalue problem

$$\begin{aligned} (C_{ff} + r_x I)^{-1} C_{fg} (C_{gg} + r_y I)^{-1} C_{gf} \hat{\alpha} &= \lambda^2 \hat{\alpha} \\ \hat{\beta} &\propto C_{gg}^{-1} C_{gf} \hat{\alpha}. \end{aligned}$$

The basis vectors for the original kernel matrices are then given as $\alpha = F^\dagger \hat{\alpha} = (F^T F)^{-1} F^T \hat{\alpha}$. The projections of the training and test acoustic features are $\hat{X} = \alpha^T K_x$ and $\hat{X}^{(\text{test})} = \alpha^T K_x^{(\text{test})}$, where $[K_x^{(\text{test})}]_{ij} = k_x(x_i, x_j^{(\text{test})})$ is the kernel evaluated at the i^{th} training and j^{th} test example.

The gram-product decomposition of Eq. 5 is the most computationally expensive step. We solve this via a block incremental SVD approach, based on [29] and detailed in [18]. This is key to making KCCA feasible for typical speech problems. Note that we can apply incremental SVD to Eq. 5 but not directly to Eq. 4, as computing the matrix of Eq. 4 is itself problematic. In our case, then, the initial dimensionality reduction of Eq. 5 is motivated not only by regularization but also (especially) by computation.

3. EXPERIMENTS

We experiment with the proposed acoustic features for phonetic recognition with 3-state monophone HMM/GMM recognizers. For most experiments, we use a subset of the University of Wisconsin X-ray Microbeam Database (XRMB) [5] of acoustic and articulatory recordings. The articulatory data consist of horizontal and vertical displacements of 8 pellets on the speaker's lips, tongue, and jaw, yielding a 16-dimensional vector at each sample. We use data from the speakers JW11, JW13, JW24, and JW30 (two male, two female). Baseline acoustic features are mean- and variance-normalized 13-dimensional mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives computed every 10ms over a 25ms window. The articulatory measurements are downsampled to match the MFCC frame rate.

The input features to CCA/KCCA are the acoustic and articulatory features concatenated over a 7-frame window around each frame, giving acoustic vectors $X \in \mathbb{R}^{273 \times N}$ and articulatory vectors $Y \in \mathbb{R}^{112 \times N}$, where N is the number of frames ($\sim 50,000$ per speaker). (Shorter windows produce worse results; longer windows may further improve results.)

For domain-independence experiments, we also use one speaker (msak0) from the MOCHA-TIMIT acoustic-articulatory database [6] and the TIMIT database [30]. The acoustic parameterization is the same. For MOCHA-TIMIT, the articulatory measurements consist of 7 tracks of pellets on the speaker's lower and upper lips; lower incisor; tongue tip, body, and dorsum; and velum. The MOCHA articulatory data is 98-dimensional after stacking over a 7-frame window.

3.1. Speaker-dependent phonetic recognition

We first explore the performance of phonetic recognition using baseline and transformed CCA/KCCA-based acoustic features in a speaker-dependent setting. In particular, we train and test CCA and KCCA transformations on data from each of the four XRMB speakers. Since the XRMB data sets are relatively small and there are several repeated utterances, a phonetic language model learned on XRMB would be too artificial; instead, we use a bigram phonetic language model learned on TIMIT data. For KCCA, we use radial basis function (RBF) kernels for both views, $k_x(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma_x^2}$ and similarly for k_y.

We use a five-fold experimental setup: We run five independent experiments, each using 60% of the utterances for learning projections and HMM/GMM parameters, 20% for tuning (development), and 20% for final testing. The error rates we report are averages over the five non-overlapping test sets. We tune all hyper-parameters (dimensionality k, regularization parameters r_x, r_y, number of Gaussians, language model penalty and scale) independently on each development set. Kernel bandwidths are fixed at $\sigma_x = 4 \times 10^6$, $\sigma_y = 2 \times 10^4$ (based on the variance of the data in each view). The intermediate dimensionality m in KCCA is fixed at 500.

As in previous work [17, 18], performance is better when concatenating the CCA-based projections with the baseline MFCCs, rather than using the CCA-based projections alone. Intuitively, there is discriminative information in the MFCCs that is not correlated with the articulatory data (e.g., voicing and nasality, which is missing in XRMB). Here we report only results with CCA and KCCA projections concatenated with the baseline MFCCs, which we refer to as "MFCCA" (MFCC+CCA) and "KMFCCA" (MFCC+KCCA).

The first four lines of Table 1 show speaker-dependent phonetic recognition results for each of the four XRMB speakers. In all cases, MFCCA improves over the baseline MFCCs by 4.5-8% absolute, and KMFCCA improves upon that by another 0.8-3.5% absolute. We also compare to unsupervised dimensionality reduction with PCA. While PCA over the 7-frame windows does improve over the baseline 39-dimensional MFCCs, the CCA-based projections still improve over PCA by several percent for all speakers.

3.2. Speaker-independence

To test the degree of speaker-independence of the CCA and KCCA projections, we next repeat the above experiments, but learn the projections on three of the XRMB speakers ("source" speakers) and test on the fourth ("target" speaker). The HMM/GMM decoder is still learned on the target speaker, so that we only measure the speaker-independence of the features and not of the statistical model.

	corpus	XRMB			
	test set	JW11	JW30	JW13	JW24
baseline	MFCC	40.5	39.0	31.3	39.0
speaker-dep	PCA	37.8	37.2	33.6	36.0
	MFCCA	35.4	34.8	26.8	31.0
	KMFCCA	31.9	33.0	26.0	30.2
cross-speaker	MFCCA	37.5	38.4	27.3	32.6
	KMFCCA	32.9	34.9	26.3	30.1

Table 1. Phonetic recognition error rates (in %) on XRMB speakers (averaged over five test sets for each speaker).

	corpus	MOCHA msak0		TIMIT	
	test set	dev	test	dev	test
baseline	MFCC	41.4	42.3	33.3	33.6
domain-dep	MFCCA	39.9	41.2	N/A	N/A
cross-dom	MFCCA	39.7	40.9	32.5	33.3

Table 2. Phonetic recognition error rates (in %) on MOCHA msak0 and the TIMIT multi-speaker dev/test sets.

The CCA/KCCA dimensionality is also tuned on the target speaker's development data; i.e., we learn an entire full-dimensional set of projections on the source speakers, and then choose which subset of the projections will be used on the new speaker. The results are shown in Table 1, averaged over the same five test sets for each speaker as before.

The phone recognition performance is very similar for the speaker-dependent and speaker-independent settings. This suggests that features learned using CCA/KCCA are speaker-independent. This is very encouraging, considering that no cross-speaker normalization or adaptation has been done. In order to make sure that we are not benefiting from the increased amount of training data in the speaker-independent cases, we randomly subsampled by a factor of three to match the training set size to that of the speaker-dependent setting.

3.3. Domain-independence

For cross-domain experiments, we choose one set of CCA directions learned on three XRMB speakers (JW11, JW24, JW30) and test them on a speaker (msak0) from the MOCHA-TIMIT corpus and on the multi-speaker full test set of TIMIT. In the case of MOCHA-TIMIT, the dialect is British English, whereas in XRMB and TIMIT it is American English. The HMM/GMM decoder is trained, and the CCA dimensionality tuned, on development data in the target domain, so that we are only testing the portability of the acoustic projections across domains. We decode with a bigram phone language model estimated from training data in these domains.

Table 2 gives the results on MOCHA-TIMIT and TIMIT. We include both development and test set results to show the effect of tuning. For MOCHA-TIMIT, we define a 5-fold experimental setup analogously to the XRMB experiments. For TIMIT, we train on the standard training set, tune on a held-out set of 50 speakers across dialect regions (similarly to [31]), and test on the full test set. For MOCHA-TIMIT, Table 2 also includes in-domain speaker-dependent results.

For MOCHA, in-domain MFCCA gives an improvement of 1.1% absolute and cross-domain (learned on XRMB) MFCCA gives an improvement of 1.4% absolute on the test set. For TIMIT, cross-domain MFCCA gives an improvement of 0.8% on the development set and 0.3% on the test set.

Interestingly, then, the effect of CCA-based features is much smaller on MOCHA-TIMIT and TIMIT than on XRMB, including the domain-dependent MOCHA-TIMIT case. We hypothesize that this is because the language model is much stronger for these corpora, dwarfing the effect of the acoustic transformations. To test this hypothesis, we also measured phonetic frame classification error rates on MOCHA-TIMIT, using a k-nearest neighbor classifier. On this task, the transformed acoustic features do improve the error rate by $5 - 6.5\%$ absolute over the baseline MFCCs (from 46.6% baseline frame error to 41.5% with domain-dependent MFCCA and 40.1% with cross-domain MFCCA). These are similar improvements to those we have seen on XRMB [18]. Therefore, on MOCHA-TIMIT there is a large gap between frame classification and phonetic recognition, lending support to our hypothesis that the language model accounts for the difference. It will therefore be interesting to consider other, more realistic corpora and tasks in the future.

4. CONCLUSION

Our results show that CCA-based acoustic features learned using articulatory measurements are useful for phonetic recognition, are largely speaker-independent, and also are domain-independent to some extent. This extends previous attempts to use multi-view learning of features that were limited to phonetic frame classification and to speaker-dependent experiments. This also takes us a step closer to the pursuit of an improved generic front-end for speech recognition using additional measurements available only at training time and only for some limited corpora. We leave for future work a more thorough comparison of CCA-based features with other unsupervised transformations besides PCA (e.g., [32]).

Our experiments thus far have not considered speaker or domain adaptation. It is plausible that the coordinate system of the most correlated subspace between acoustics and articulation may differ across speakers and domains, and that a combination of multi-view learning and unsupervised adaptation may provide additional gains. We have also not yet explored the full KCCA tuning space, including alternative kernels and regularizers [33], nor alternative input features besides MFCCs. Finally, additional future work will study whether the gains apply to more complex tasks such as word recognition, as well as extensions based on views other than articulatory tracks (video [34], EMG [35], MRI [36], ultrasonic signals [37], etc.) and based on semi-supervised extensions combining multiple signal views as well as labels [38].

5. REFERENCES

[1] R. Haeb-Umbach and H. Ney, “Linear discriminant analysis for improved large vocabulary continuous speech recognition,” in *ICASSP*, 1992.

[2] N. Kumar and A. G. Andreou, “Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition,” *Speech Comm.*, vol. 26, no. 4, pp. 283–297, 1998.

[3] H. Hermansky *et al.*, “Tandem connectionist feature extraction for conventional HMM systems,” in *ICASSP*, 2000.

[4] S. King *et al.*, “Speech production knowledge in automatic speech recognition,” *Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[5] J. R. Westbury, *X-ray microbeam speech production database user’s handbook*, Waisman Center on Mental Retardation & Human Development, U. Wisconsin, Madison, WI, version 1.0 edition, June 1994.

[6] A. Wrench, “A new resource for production modeling in speech technology,” in *Workshop on Innovations in Speech Processing*, 2001.

[7] K. Richmond, P. Hoole, and S. King, “Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus,” in *Interspeech*, 2011.

[8] J. Frankel and S. King, “ASR - articulatory speech recognition,” in *Eurospeech*, 2001.

[9] A. Wrench and K. Richmond, “Continuous speech recognition using articulatory data,” in *ICSLP*, 2000.

[10] T. Toda, A. W. Black, and K. Tokuda, “Acoustic-to-articulatory inversion mapping with Gaussian mixture model,” in *ICSLP*, 2004.

[11] B. Uria *et al.*, “Deep architectures for articulatory inversion,” in *Interspeech*, 2012.

[12] K. Markov *et al.*, “Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework,” *Speech Communication*, vol. 48, pp. 161–175, 2006.

[13] L. Deng *et al.*, “Production models as a structural basis for automatic speech recognition,” *Speech Comm.*, vol. 33, pp. 93–111, 1997.

[14] M. Richardson, J. Bilmes, and C. Diorio, “Hidden-articulator Markov models for speech recognition,” *Speech Comm.*, vol. 41, no. 2–3, pp. 511–529, 2003.

[15] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, “Gesture-based dynamic Bayesian network for noise robust speech recognition,” in *ICASSP*, 2011.

[16] K. Livescu *et al.*, “Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop,” in *ICASSP*, 2007.

[17] S. Bharadwaj *et al.*, “Multiview acoustic feature learning using articulatory measurements,” in *Intl. Workshop on Stat. Machine Learning for Speech Processing*, 2012.

[18] R. Arora and K. Livescu, “Kernel CCA for multi-view learning of acoustic features using articulatory measurements,” in *Symp. on Machine Learning in Speech and Language Processing (MLSLP)*, 2012.

[19] K. Sridharan and S. Kakade, “An information theoretic framework for multi-view learning,” in *Conference on Learning Theory*, 2008.

[20] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[21] D. R. Hardoon *et al.*, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[22] K. Chaudhuri *et al.*, “Multi-view clustering via canonical correlation analysis,” in *ICML*, 2009.

[23] K. Livescu and M. Stoehr, “Multi-view learning of acoustic features for speaker recognition,” in *ASRU*, 2009.

[24] K. Choukri and G. Chollet, “Adaptation of automatic speech recognizers to new speakers using canonical correlation analysis techniques,” *Speech Comm.*, vol. 1, pp. 95–107, 1986.

[25] F. Rudzicz, “Adaptive kernel canonical correlation analysis for estimation of task dynamics from acoustics,” in *ICASSP*, 2010.

[26] T. Kato *et al.*, “An analysis of articulatory-acoustic data based on articulatory strokes,” in *ICASSP*, 2009.

[27] D. Foster, S. Kakade, and T. Zhang, “Multi-view dimensionality reduction via canonical correlation analysis,” Tech. Rep. TTIC-TR-2008-4, TTI-Chicago, 2008.

[28] K. Fukumizu *et al.*, “Statistical consistency of Kernel Canonical Correlation Analysis,” *Journal of Machine Learning Research*, vol. 8, pp. 361–383, 2007.

[29] M. Brand, “Incremental singular value decomposition of uncertain data with missing values,” in *Eur. Conf. on Comp. Vision*, 2002.

[30] J. S. Garofolo *et al.*, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM,” *NASA STI/Recon Technical Report N*, 1993.

[31] A. K. Halberstadt, *Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition*, Ph.D. thesis, MIT, 1998.

[32] A. Jansen, S. Thomas, and H. Hermansky, “Intrinsic spectral analysis for zero and high resource speech recognition,” in *Interspeech*, 2012.

[33] D. R. Hardoon and J. Shawe-Taylor, “Sparse canonical correlation analysis,” *Machine Learning*, vol. 83, pp. 331–353, 2011.

[34] G. Potamianos and C. Neti, “Audio-visual speech recognition in challenging environments,” in *Eurospeech*, 2003.

[35] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, “Towards continuous speech recognition using surface electromyography,” in *Interspeech*, 2006.

[36] S. Narayanan *et al.*, “A multimodal real-time MRI articulatory corpus for speech research,” in *Interspeech*, 2011.

[37] K. Livescu, B. Zhu, and J. Glass, “On the phonetic information in ultrasonic microphone signals,” in *ICASSP*, 2009.

[38] S. Yu, B. De Moor, and Y. Moreau, “Learning with heterogenous data sets by weighted multiple kernel canonical correlation analysis,” in *Machine Learning for Signal Processing (MLSP)*, 2007.