MULTI-VIEW LEARNING WITH SUPERVISION FOR TRANSFORMED BOTTLENECK FEATURES

Raman Arora and Karen Livescu

TTI-Chicago, Chicago, IL 60637

arora@ttic.edu, klivescu@ttic.edu

ABSTRACT

Previous work has shown that acoustic features can be improved by unsupervised learning of transformations based on canonical correlation analysis (CCA) using articulatory measurements that are available at training time. In this paper, we investigate whether this second view (articulatory data) still helps even when labels are also available at training time. We begin with strong baseline bottleneck features, which can be learned when the training set is phonetically labeled. We then compare several options for learning transformations of the bottleneck features in the presence of both articulatory measurements and phonetic labels for the training data. The methods compared include combinations of LDA and CCA, as well as a three-view extension of CCA that simultaneously uses the labels and articulatory measurements as additional views. Phonetic recognition experiments on data from the University of Wisconsin X-ray microbeam database show that the learned features improve performance over using either just the labels or just the articulatory measurements for learning acoustic transformations.

Index Terms— multi-view learning, canonical correlation analysis, articulatory measurements, bottleneck features, supervised transformation learning

1. INTRODUCTION

One of the main recent improvements to automatic speech recognition has come from improving the acoustic features by learning transformations of basic features. A typical approach is to construct high-dimensional features, by concatenating multiple consecutive frames of basic features (e.g., mel-frequency cepstral coefficients, perceptual linear prediction coefficients) around the current frame, and then to reduce dimensionality using a transformation that is learned from data. Such learned transformations can include:

- Unsupervised linear transformations, such as principal components analysis (PCA)
- Unsupervised nonlinear transformations, such as graph-based methods [1]
- Supervised linear transformations, such as linear discriminant analysis (LDA) and its extensions [2, 3]
- Supervised nonlinear transformations, typically using neural networks with phonetic (or state) labels as outputs, such as tandem [4] and bottleneck [5] processing

A research question that has recently been pursued is whether it is possible to learn better transformations if we have access to another *view* of the speech data at training time, in particular articulatory measurements. Articulatory measurements clearly help if they are available at test time [6, 7], and may also help in the more realistic scenario where they are only available at training time and not test time [8]. We have found that such an additional view can indeed help learn improved acoustic feature transformations, using a multi-view learning approach based on canonical correlation analysis (CCA) [9, 10, 11]. Encouragingly, the improvements seem to hold for speakers for whom no articulatory measurements are available even at training time [11].

Our previous work on CCA-based acoustic feature transformation learning has been in the unsupervised setting. There are certain virtues to this setting, but frame labels are often available, or relatively high-quality ones can be obtained via forced transcription. In such cases much better acoustic features can be obtained, in particular ones based on neural networks as in the tandem [4] and bottleneck [5] approaches. In this paper, we ask whether in this supervised setting, we can start from strong supervised features (in our case bottleneck features) and still obtain an improved feature transformation using the additional information from articulatory measurements *and* labels at training time.

We consider several ways of combining a second view with labels, using either combinations of CCA and LDA or three-view extensions of CCA that consider the labels to be a third view. Fig. 1 gives a pictorial overview of the approach. We present phonetic recognition experiments on data from the University of Wisconsin X-ray microbeam database (XRMB) [12], comparing the techniques and showing that indeed the new transformations improve over both unsupervised CCA or supervised transformations alone.

This research was supported by NSF grants IIS-0905633 and IIS-1321015. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

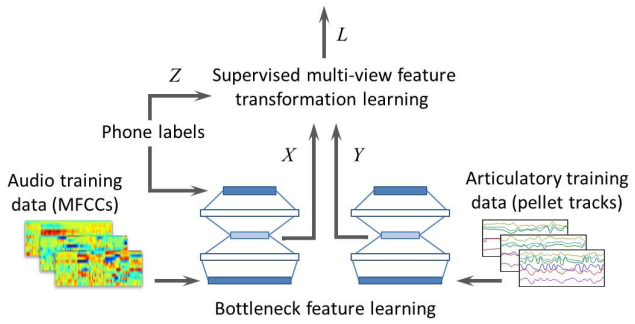

Fig. 1. Pictorial overview of our approach and notation.

2. METHODS

In this section we review unsupervised multi-view feature transformation learning via CCA, as well as supervised single-view transformation learning via LDA, and then describe several ways of using both labels and multiple views. Here we only consider linear transformations, although the techniques have nonlinear extensions. Our training data are N samples of random vectors of acoustic features, $X \in \mathbb{R}^{d_x \times N}$, articulatory measurement features $Y \in \mathbb{R}^{d_y \times N}$, and labels $Z \in \mathcal{Z}$, where the definition of $\mathcal{Z}$ depends on how the labels will be used (see below). The training set is then $\{x_i, y_i, z_i\}_{i=1}^N$, where each tuple (x_i, y_i, z_i) represents features computed over one frame of simultaneously recorded acoustics (x_i) and articulation (y_i) and the corresponding phone label (z_i). To simplify notation we assume that the data have been mean-centered. Our goal is to find a transformation matrix $L \in \mathbb{R}^{d_x \times m}, m \leq d_x$, such that the transformed acoustic features $L^\top X$ improve recognition performance. Fig. 1 shows our overall setup and notation.

Unsupervised multi-view learning via CCA: In previous work in the unsupervised setting (i.e., given only samples of raw features X and Y), we have taken an approach based on canonical correlation analysis (CCA) [10] and its nonlinear extensions [11, 13]. We review this approach for completeness. CCA is a technique for learning maximally correlated linear projections of data in two views [14, 15]. We give the formulation of Golub and Zha [16] (instead of the common statistical formulation [14]) as it clarifies connections with other methods: Given the pair of data matrices $X \in \mathbb{R}^{d_x \times N}$ and $Y \in \mathbb{R}^{d_y \times N}$, the CCA projection vectors are columns of the matrices $\hat{U} \in \mathbb{R}^{d_x \times m}$ and $\hat{V} \in \mathbb{R}^{d_y \times m}$ that solve:

$$\begin{aligned} \underset{U,V}{\text{minimize}} \quad & \left\| U^\top X - V^\top Y \right\|_F^2 \\ \text{subject to} \quad & \frac{1}{N} U^\top X X^\top U = I, \frac{1}{N} V^\top X X^\top V = I, \\ & \frac{1}{N} U^\top X Y^\top V = I \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This problem is equivalent to finding maximally correlated pairs of projections, subject to uncorrelatedness of subsequent projections. We will use the following shorthand for Problem (1):

$$\left\{ \hat{U}, \hat{V} \right\} = \arg\max_{U,V} \rho(U^\top X, V^\top Y). \tag{2}$$

where ρ represents correlation (in this case it should be thought of as multidimensional correlation as above). The solution of Eq. (1) is straightforward [15]: The matrix $\hat{U}$ consists of the top m eigenvectors of $\hat{C}_{xx}^{-1} \hat{C}_{xy} \hat{C}_{yy}^{-1} \hat{C}_{yx}$, where $\hat{C}_{xx} = \frac{1}{N} X X^\top, \hat{C}_{yy} = \frac{1}{N} Y Y^\top$ are the sample auto-covariance matrices and $\hat{C}_{xy} = \frac{1}{N} X Y^\top$ is the sample cross-covariance matrix. The matrix $\hat{V}$ is composed of columns $v_k \propto \hat{C}_{yy}^{-1} \hat{C}_{yx} u_k$; in our case, we discard $\hat{V}$ since we only have access to the acoustic view at test time. In practice, in order to avoid ill-conditioned auto-covariance matrices and mitigate the effects of noise in the data, we regularize by adding a term $r_x I$ to the auto-covariance estimate $\hat{C}_{xx}$ and $r_y I$ to $\hat{C}_{yy}$.

The largest possible number of learned projections is $\min(d_x, d_y)$; in order to reduce dimensionality, we retain only the top $m < \min(d_x, d_y)$ projection vectors, and these form the desired final transformation matrix $L = \hat{U} = [u_1 \ldots u_m]$. Nonlinear extensions of CCA can be derived by replacing the linear projections with nonlinear functions $f(X), g(Y)$ defined via kernels [17, 15] or neural networks [18, 13].

A common motivation for the use of CCA is that if the two views are largely uncorrelated conditioned on some class of interest (here, the phonetic class), then the dimensions that are correlated should be discriminative. In our case, a second motivation is that we know the second view (articulation) to be useful if available at test time, but this is an unrealistic setting, and predicting articulation from acoustics is very challenging and perhaps not necessary; finding good correlated projections may be sufficient.

Supervised single-view learning via LDA: When we have labels and a single (acoustic) view, it is common to use linear discriminant analysis (LDA) or its extensions [2, 3]. LDA learns a discriminative transformation matrix L that maps feature vectors from the same class to nearby vectors while maximally separating the projected class means. It is easy to check that LDA is a special case of CCA where the second view is the class label represented in the following way (a "one-hot" representation): If the i^{th} training sample belongs to class c, where $1 \leq c \leq C$, then $\mathcal{Z} = \{0, 1\}^{C \times N}$ and the k^{th} dimension of Z is 1 if $k = c$ and 0 otherwise:

$$[z_i]_k = \begin{cases} 1 & \text{if } k = c, \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Supervised single-view learning via neural networks: In the single-view supervised setting, neural networks are often used to learn nonlinear transformations. The transformed features are outputs of either the network's final layer (as in tandem systems [4]) or a narrow hidden layer (bottleneck features [5]). In this work, we use bottleneck features and consider these to be our input X for all supervised methods.

2.1. Supervised multi-view learning

Multi-view feature learning is typically used as an unsupervised technique, but when training labels are also available, we can use them to learn features that are more discriminative. Why might the second view (here, articulation) be useful even in the presence of labels? One reason is that this second view may "guide" the learning into a good part of the space when directly finding a discriminative subspace is difficult, for example by identifying and removing noisy dimensions. Another reason is that the labels may not be for the target task, but rather for a proxy task; this is often the case in acoustic feature learning, where phonetic frame labels are used but the ultimate task is continuous phone/word recognition.

We next describe several ways of combining labels with the second articulatory view for feature learning. We assume the labels are represented as vectors in $\mathbb{R}^C$ using the "one-hot" representation of Eq. (3). There are a number of techniques that have been suggested for this setting, including multi-view extensions of LDA [19, 20], labeled extensions of CCA [21], and generalizations that subsume both of these [22]. In this work we select a few approaches that are straightforward to solve via one or two eigenproblems.

2.1.1. CCA on bottleneck features

The first way we combine two views with labels is to simply apply CCA to the learned bottleneck features. The remaining methods below further use the labels in a second stage.

2.1.2. CCA with concatenated labels and articulatory data

One straightforward approach is to consider the labels to be part of the second view, by appending them to the articulatory data and then using CCA. The transformation matrix is then given by solving the modified CCA problem

$$\left\{\hat{U}, \hat{V}\right\} = \arg\max_{U,V} \ \rho\left(U^\top X, V^\top \left[\begin{array}{c} Y \\ Z \end{array}\right]\right) \qquad (4)$$

and the final acoustic feature transformation is $L = \hat{U}$.

2.1.3. Concatenated LDA and CCA features

We can also consider solving two multi-view problems, one with articulatory data as the second view and another with the labels as the second view, and concatenating the results; i.e., concatenating CCA and LDA features. Let U, W denote transformations learned with CCA and LDA, respectively:

$$\begin{aligned} \left\{\hat{U}, \hat{V}\right\} &= \arg\max_{U,V} \ \rho\left(U^\top X, V^\top Y\right) \qquad (5) \\ \left\{\hat{W}, \hat{T}\right\} &= \arg\max_{W,T} \ \rho\left(W^\top X, T^\top Z\right). \end{aligned}$$

Our final feature transformation matrix is then $L = \left[\hat{U}\ \hat{W}\right]$.

2.1.4. LDA on CCA features

Another option is to first learn CCA-based features, and then to learn a supervised transformation of those features via LDA. The CCA step could be done on either raw features (unsupervised) or on the supervised bottleneck features as in Sec. 2.1.1; here we only consider the latter. The final transformation is given by

$$\left\{\hat{W}, \hat{T}\right\} = \arg\max_{W,T} \ \rho\left(W^\top \hat{U}^\top X, T^\top Z\right), \qquad (6)$$

where $\hat{U}$ is given in (5). The final transformation is $L = \hat{U}\hat{W}$.

2.1.5. Many-view learning with generalized CCA (GCCA)

There are several generalizations of CCA for any number of views $J \geq 2$ [23, 24, 25]. We follow the formulation introduced by Carroll [24], which seeks a common latent representation ("group configuration") $G \in \mathbb{R}^{m \times N}$ and view-specific transformation matrices $\hat{U}_j \in \mathbb{R}^{d_j \times m}, 1 \leq j \leq J$, that solve

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{J} \left\| G - U_j^\top X_j \right\|_F^2, \\ \text{subject to} \quad & GG^\top = I \end{aligned} \qquad (7)$$

where $X_j \in \mathbb{R}^{d_j \times N}$ is the feature matrix for the j^{th} view. In our case $J = 3, X_1 = X, X_2 = Y, X_3 = Z$, and the final transformation matrix is $L = \hat{U}_1$. G is given as the solution to the eigenvalue problem [25]:

$$\left(\sum_{j=1}^{J} X_j^\top \left(X_j X_j^\top\right)^{-1} X_j\right) G^\top = G^\top \Lambda, \qquad (8)$$

and the transformation matrices $\hat{U}_j$ are given as

$$\hat{U}_j = \left(X_j X_j^\top\right)^{-1} X_j G^\top. \qquad (9)$$

As before, we implement a regularized variant by adding $r_j I$ to each autocovariance term $X_j X_j^\top$ in Eqs. (8, 9).

3. EXPERIMENTS

We test the proposed features for phonetic recognition on a subset of the University of Wisconsin X-ray Microbeam Database (XRMB) [12] of acoustic and articulatory recordings. The setup is the same as in previous related work [11]. Baseline acoustic features are mean- and variance-normalized 13-dimensional mel-frequency cepstral coefficients (MFCCs) and their 1st and 2nd derivatives. The articulatory data are horizontal and vertical displacements of 8 pellets on the speaker's lips, tongue, and jaw, yielding a 16-dimensional vector at each sample, downsampled to the MFCC frame rate. We use data from two randomly chosen male speakers (JW11, JW24) and two female speakers (JW13, JW30). The input features to the two bottleneck networks are the acoustic and articulatory features concatenated over a 7-frame window around each frame, giving 273-dimensional acoustic inputs and 112-dimensional articulatory inputs.

We compare baseline and transformed acoustic features in a speaker-dependent setting. The recognizers use 3-state monophone HMM/GMMs with a TIMIT bigram language model (LM); an XRMB LM would be too biased due to XRMB's limited utterance inventory. We use a five-fold setup

	Speaker	JW11	JW30	JW13	JW24	Average
unsupervised baseline	MFCC	32.5	34.5	26.1	30.7	31.0
unsupervised multi-view	*CCA* (MFCC, Artic)	31.4	33.0	25.9	29.4	29.9
supervised baselines	*LDA* (MFCC)	31.8	31.9	25.6	30.8	30.0
	PCA (BN_x)	33.0	29.8	23.4	27.7	28.5
	LDA (BN_x)	28.1	29.6	23.8	27.8	27.3
supervised multi-view	Sec. 2.1.1 *CCA* (BN_x, BN_y)	27.2	29.1	22.5	27.9	26.7
	Sec. 2.1.2 *CCA* (BN_x, $BN_y \oplus$ Lab)	26.9	29.4	22.8	27.4	26.6
	Sec. 2.1.3 *CCA* (BN_x, BN_y) $\oplus$ *LDA* (BN_x)	26.5	30.4	23.2	27.9	27.0
	Sec. 2.1.4 *LDA* (*CCA* (BN_x, BN_y))	26.4	29.1	**21.9**	**26.9**	26.1
	Sec. 2.1.5 *GCCA* (BN_x, BN_y, Lab)	**25.7**	**28.9**	22.3	**26.9**	**25.9**

Table 1. Revised phonetic error rates (%) averaged over five test sets for each speaker, and average over the speakers. In all cases except unsupervised baselines, MFCCs are appended to the indicated features. Notation: Artic = raw articulatory features; BN_x = acoustic bottleneck features; BN_y = artic. bottleneck features; Lab = labels in "one-hot" representation; "*METHOD* (features)" denotes application of *METHOD* to the features in parentheses; $\oplus$ denotes concatenation. Example: *CCA* (BN_x, $BN_y \oplus$ Lab) denotes CCA where one view is acoustic bottleneck features and the second view is articulatory bottleneck features appended to labels. See the indicated sections for further details.

with five separate experiments for each speaker, each using 60% of the data for learning projections and HMM/GMM parameters, 20% for development, and 20% for testing. We report average error rates over the five non-overlapping test sets. For baseline bottleneck experiments, we use a bottleneck network topology of 273–1745–35–1745–39, followed by either a PCA rotation or LDA rotation and optional dimensionality reduction, following the setup of [5]. For multi-view experiments, a bottleneck layer of size 150 is used in order to allow for more freedom in dimensionality reduction experiments. We use rectified linear units and pre-training with de-noising auto-encoders (DAE) and dropout. Following [26], we initialize the weights at the i^{th} DAE layer uniformly randomly in the interval $\left[-\sqrt{\frac{6}{w_{i-1}+w_i}}, \sqrt{\frac{6}{w_{i-1}+w_i}}\right]$, where w_i is the width of the i^{th} layer. The learning rate was chosen using the adaptive learning rate method ADADELTA [27]. We used the Deepmat toolkit for learning bottleneck features [28]. We tune hyper-parameters (dimensionalities, regularization parameters r_x, r_y, number of Gaussians, LM penalty/scale) independently on each development set. For GCCA, we sub-sample the data to $10,000$ frames for improved speed (a possible future improvement is to use incremental matrix factorization with all of the data, as in [11]). As in previous work, performance is better when concatenating the learned features with the baseline MFCCs; all of the given results are therefore with concatenated features. We emphasize that the HMM/GMMs are trained and tested on acoustic data only; the articulatory data is not used after the transformation learning step.

Table 1 shows the phonetic error rates for the four speakers, as well as averages over all speakers. For comparison with earlier unsupervised multi-view work, we include results for CCA applied to raw MFCCs and articulatory features[1]. As might be expected, the best supervised baselines improve over unsupervised multi-view techniques. Supervised multi-view techniques yield an additional improvement over supervised single-view techniques. Average improvements over the best supervised baseline (LDA on bottleneck features) range from 0.7% to 2.4% absolute (2-9% relative). While the supervised multi-view methods all perform similarly, the most successful method on average (as well as on most speakers) is many-view GCCA with labels as a third view.

[1]Baseline MFCC and CCA results are improved over our earlier results in [11], due to improved tuning. All systems here are tuned in the same way.

4. CONCLUSION

Our results indicate that acoustic features learned using articulatory measurements and phonetic labels, via supervised multi-view learning, are useful for phonetic recognition. Not surprisingly, the use of labels to learn bottleneck features improves over previous unsupervised results. What is less self-evident is that a second data view (here, articulatory measurements) should help on top of the supervised bottleneck features. The fact that multi-view learning helps here even in the presence of labels suggests that there is indeed something to be learned from such an auxiliary view, perhaps by making the discriminative feature learning easier, or perhaps because the discriminative problem (frame classification) is not precisely the same as our ultimate goal (continuous recognition).

Previous work has shown that in the unsupervised multi-view case, nonlinear transformations provide further improvements over linear ones, and the features generalize to new speakers for whom no articulatory data are available even at training time [11]. These are natural extensions for our work here with the supervised case. Other directions for future work include additional supervised multi-view approaches [22, 23, 29], as well as the use of additional views besides articulatory tracks. To our knowledge only video has been used with speech in a similar multi-view setting [30, 31], but any of the other types of measurements being explored for speech (MRI [32, 33], ultrasonic sensors [34, 35], tongue ultrasound [36], muscle EMG [37]) could be good candidates.

5. REFERENCES

[1] A. Jansen *et al.*, "Intrinsic spectral analysis for zero and high resource speech recognition," in *Interspeech*, 2012.

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

[3] 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.

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

[5] F. Grezl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *ICASSP*, 2008.

[6] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *ICSLP*, 2000.

[7] J. Frankel and S. King, "ASR - articulatory speech recognition," in *Eurospeech*, 2001.

[8] L. Badino *et al.*, "Deep-level acoustic-to-articulatory mapping for DBN-HMM based phone recognition," in *SLT*, 2012.

[9] S. Bharadwaj *et al.*, "Multiview acoustic feature learning using articulatory measurements," in *Intl. Workshop on Stat. Machine Learning for Speech Processing (IWSML)*, 2012.

[10] 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.

[11] R. Arora and K. Livescu, "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," in *ICASSP*, 2013.

[12] 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.

[13] G. Andrew *et al.*, "Deep canonical correlation analysis," in *ICML*, 2013.

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

[15] 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.

[16] G. H. Golub and H. Zha, "The canonical correlations of matrix pairs and their numerical computation," *IMA Volumes in Mathematics and its Applications*, 1995.

[17] S. Akaho, "A kernel method for canonical correlation analysis," in *Intl. Meeting of the Psychometric Society*, 2001.

[18] H. Asoh and O. Takechi, "An approximation of nonlinear canonical correlation analysis by multilayer perceptrons," in *ICANN*, 1994.

[19] T. Diethe *et al.*, "Constructing nonlinear discriminants from multiple data views," in *ECML PKDD*, 2010.

[20] Q. Chen and S. Sun, "Hierarchical multi-view Fisher discriminant analysis," in *ICONIP*, 2009.

[21] L. Sun *et al.*, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 194–200, 2011.

[22] A. Sharma *et al.*, "Generalized multiview analysis: A discriminative latent space," in *CVPR*, 2012.

[23] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[24] J. D. Carroll, "Generalization of canonical correlation analysis to three or more sets of variables," in *76th Annual Convention of the American Psychological Association*, 1968.

[25] M. van de Velden and T. H. A. Bijmolt, "Generalized canonical correlation analysis of matrices with missing rows: a simulation study," *Psychometrika*, vol. 71, no. 2, pp. 323–331, 2006.

[26] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[27] Matthew D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.

[28] Kyunghyun Cho, "Deepmat toolkit," https://github.com/kyunghyuncho/deepmat, 2013.

[29] S. Yu *et al.*, "Learning with heterogenous data sets by weighted multiple kernel canonical correlation analysis," in *MLSP*, 2007.

[30] M. E. Sargin *et al.*, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE. Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

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

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

[33] K. Richmond *et al.*, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Interspeech*, 2011.

[34] K. Kalgaonkar and B. Raj, "Ultrasonic Doppler sensor for speaker recognition," in *ICASSP*, 2008.

[35] K. Livescu *et al.*, "On the phonetic information in ultrasonic microphone signals," in *ICASSP*, 2009.

[36] T. Hueber *et al.*, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Comm.*, vol. 52, no. 4, pp. 288–300, 2010.

[37] S.-C. Jou *et al.*, "Towards continuous speech recognition using surface electromyography," in *Interspeech*, 2006.