*INTERSPEECH 2017*
August 20–24, 2017, Stockholm, Sweden

# Learning weakly supervised multimodal phoneme embeddings

*Rahma Chaabouni*[1], *Ewan Dunbar*[1], *Neil Zeghidour*[1,2], *Emmanuel Dupoux*[1]

[1]Département d'Etudes Cognitives, Ecole Normale Supérieure, Ecole des Hautes Etudes en Sciences Sociales, PSL Research University, Centre National de la Recherche Scientifique, Paris, France.
[2]Facebook AI Research, Paris

chaabounirahma,emmanuel.dupoux, neil.zeghidour@gmail.com,emd@umd.edu

## Abstract

Recent works have explored deep architectures for learning multimodal speech representation (e.g. audio and images, articulation and audio) in a supervised way. Here we investigate the role of combining different speech modalities, i.e. audio and visual information representing the lips' movements, in a weakly supervised way using Siamese networks and lexical same-different side information. In particular, we ask whether one modality can benefit from the other to provide a richer representation for phone recognition in a weakly supervised setting. We introduce mono-task and multi-task methods for merging speech and visual modalities for phone recognition. The mono-task learning consists in applying a Siamese network on the concatenation of the two modalities, while the multi-task learning receives several different combinations of modalities at train time. We show that multi-task learning enhances discriminability for visual and multimodal inputs while minimally impacting auditory inputs. Furthermore, we present a qualitative analysis of the obtained phone embeddings, and show that cross-modal visual input can improve the discriminability of phonological features which are visually discernable (rounding, open/close, labial place of articulation), resulting in representations that are closer to abstract linguistic features than those based on audio only.

**Index Terms**: language acquisition, multimodal learning, unit of sound representation, weakly supervised learning, speech recognition, Siamese network, ABX

## 1. Introduction

The ability of many people, hearing and non-hearing, to lip read, demonstrates that speech perception is not only a purely auditory skill. Audio-visual integration is illustrated clearly by the McGurk effect [1]: the lip movements corresponding to [g], presented together with audio corresponding to a [b], is perceived as an intermediate sound, identified as [d], by many subjects. Such interactions between modalities have been documented in 5-month-old infants [2, 3]. The visual channel for speech is poorer than the auditory channel, but provides information which can be complementary, especially regarding place of articulation [4] (for example, between the coronal and labial consonants [d] and [b]).

Previous research has investigated the use of cross-modal information for speech recognition and has focused on systems trained with supervised learning (phoneme labels). Here, we investigate the case of weakly supervised learning, which is more appropriate for the modeling of infants' language acquisition. In particular, we will use a Siamese DNN architecture which feeds on word-level side information (the fact that two words are the same or different: [5, 6, 7, 8]), which demonstrably can be discovered automatically from continuous speech using spoken term discovery ([7, 9]). In this paper, we use gold word-level annotations on the Blue-Lips corpus and focus on the combination of modalities in *mono-task* and *multi-task* settings.

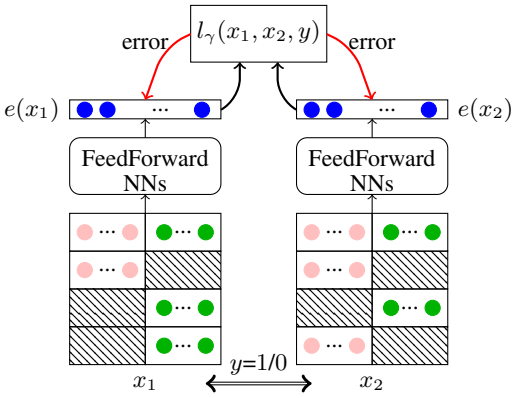


Figure 1: *Multi-task setting. $x_1$ and $x_2$ are the training inputs and $y$ the output (1 if the inputs are the same words, and 0 if not); the pink circles are the acoustic features and the green circles the visual ones. The shaded area are for zeroed inputs.*

In mono-task settings, we train a Siamese network (see below) on only one type of input (audio only, visual only, or the concatenation of the two modalities). The multi-task setting uses a form of data augmentation in which we selectively knock out one of the modalities by setting all of the input features to zero. We include the four combinations shown in Figure 1, all of which are presented to the network during training, with equal probability. In the first three cases, the input to the two branches of the Siamese network is the same; the fourth case can be thought of as a lip-reading task, in which the network is presented with different modalities in each branch.

In the following sections, we first discuss related work. Then, we describe our model and the different evaluation measures. Finally, we report experimental results and conclude.

## 2. Related work

Classical audio-visual ASR systems use an audio-visual corpus to build complex supervised classifiers able to have an efficient phoneme representation [10, 11]. While this approach benefits from a rich phoneme representation that leads to good speech recognition, it needs thousands of hours of annotated speech, which is implausible as a model of language acquisition. Some other studies involve both geometry-based and appearance-based features to build less complex models [12] but rely on an important upfront knowledge of optimal features. A more recent architecture is the *WLAS* [13], which yields state-



http://dx.doi.org/10.21437/Interspeech.2017-1689

of-the-art performance on lip reading. However, it focuses primarily on the lip reading task (hence only on the acoustic contribution to the visual inputs), and is in a supervised setting.

Our study deals with audio-visual ASR in a weakly supervised setting and evaluates the contribution of each modality in speech recognition. One of the most closely related works is [14], which learns an audio-visual speech representation in an unsupervised setting, but proposes to train a supervised classifier to evaluate the learned representation. As discussed in [15], the supervised classification performance obtained on features is not a reliable indicator of the performance of unsupervised algorithms. Supervised learning may improve certain weaknesses in the features, such as poor scaling or noisy channels, which would present major obstacles in an unsupervised setting when using clustering algorithms. Rather, in our study, we evaluate different properties of the representation: its phonetic discriminability and its parallelism. Those measures will be introduced in the following sections.

Another relevant model is the DCCA [16, 17] which can learn complex non-linear transformations of two data views to give a highly correlated embedding representation. However, maximizing the correlation in two views, in our case, the acoustic and visual views of the speech, would not necessarily lead to an efficient phoneme representation. Indeed, the acoustic signal can fully explain phonetic class identity, while it is underspecified in the visual modality. Hence, maximizing the correlation between these two modalities may lead to a loss of information for phonemes that share the same visual correlates.

We use an ABnet type architecture [6], a particular type of Siamese network that allows learning phonetic-level embeddings from word-level annotations. Such an architecture previously showed good performance [7] in the context of the Zero Resource Speech Challenge 2015 [18].

# 3. Methods

## 3.1. Dataset

We use the Blue-Lips database [19], an audio-visual speech corpus composed of 238 French sentences read by 16 speakers of Hexagonal French. Each speaker's recordings run around 20 minutes.

We represent the audio signal with 40 MFCCs computed from 40 mel-scale filterbanks sampled at 100 frames per second, resulting in a 40 dimensional vector. These 40 coefficients are the first 40 cepstral coefficients and do not contain delta/delta-delta nor energy features.

We use two kinds of visual features. First, we use dimensionality-reduced pixels from the region of the image corresponding to the mouth, identified using a Haar cascade classifier [20] to detect the mouth at the first frame and the mean-shift algorithm [21] to track the mouth throughout the rest of the video. We match the video and audio frame rates by converting the video from 25 to 100 fps using ffmpeg.[1] This results in sequences of four identical frames for each original frame of video. The pixels for the mouth are spatially down-sampled to $30 \times 50$ pixels then whitened and reduced to 40 dimensions using PCA. We found that this did not reduce performance, consistent with [10]. Second, we concatenate these video features with lip landmarks, extracted using the active apperance model [22], a facial alignment algorithm which gives the shape of the mouth in 20 two-dimensional points. To match the audio frame rate here we apply cubic interpolation.

[1] https://ffmpeg.org/ffmpeg.html

Finally, we apply mean-variance normalization to the features, for both modalities.

## 3.2. Siamese network and ABnet

A Siamese network is an architecture that contains two identical copies of the same network, so that the two subnetworks have the same configuration and share the same parameters.

In our experiments, we use ABnet, a particular Siamese network architecture. The model is represented in Figure 1. It uses pairs of words to learn a representation of phones. The input to the network during training consists of pairs of stacked frames of MFCC features $x_1$ and $x_2$ and a label $y \in \{0, 1\}$, where $y = 1$ if $x_1$ and $x_2$ represent the same word, and $y = 0$ otherwise. $x_1$ and $x_2$ represent stacks of frames that are in correspondence across the two words: for same-word pairs, this corresponence is the result of an alignment using dynamic time warping (DTW) [23], and, for different-word pairs, an alignment along the diagonal.

The idea of this architecture is that, given an abstract notion of similarity $D$, we can learn a representation in which the distance between the two embedding-space representations $e(x_1)$ and $e(x_2)$ reflects the similarity between the inputs $x_1$ and $x_2$: we want $D(e(x_1), e(x_2))$ to be small if $x_1$ and $x_2$ represent the same word, and large otherwise. In order to achieve this goal, the ABnet is trained with the margin cosine loss function:

$$l_\gamma(x_1, x_2, y) = \begin{cases} -cos(e(x_1), e(x_2)) & \text{if } y = 1 \\ max(0, cos(e(x_1), e(x_2)) - \gamma) & \text{otherwise} \end{cases}$$

where $\gamma$ is the margin.

In our experiments, we took a margin of $0.5$. Each of the inputs $x_1$ and $x_2$ consists of 7 stacked frames of one of the modalities, or of the concatenation of the modalities. Each subnetwork contains $5$ hidden layers of $1000$ units with ReLU activations, and two output embeddings each of $39$ dimensions.

## 3.3. Evaluation

We evaluate two aspects of the speech representation: its ability to discriminate between phonemes, and its internal structure, which we probe to see whether phonological features are clearly coded.

### *3.3.1. ABX task*

To evaluate the phonetic discriminability of our embeddings, we use an ABX discrimination task [24, 25]. The task consists in presenting three stimuli $A$, $B$ and $X$, with $A$ and $B$ belonging to two different phonetic categories, and $X$ belonging to one of those categories (concretely, always $A$). Given a measure of divergence $D$ (not necessarily a proper distance metric), if $D(A, X) < D(B, X)$, then the score is 1 (success), and otherwise the score is 0 (failure).

In our experiments, $A$ and $B$ are minimal pairs of triphones: they are pairs of sounds composed each of three phonemes and differing only in their central phoneme (for example, *beg,* /bɛg/ and *bag,* /bæg/, although the triphones need not be words). The first task is a within speaker task (WST), where the three stimuli are uttered by the same speaker. It measures how the relative distance between speech utterances in the embedding space correlates with their phonetic content. The second task is an across speaker task (AST), where $A$ and $B$ are uttered by the same speaker and $X$ by a different speaker. This task is harder than the previous one since it requires embeddings to mostly reflect

Table 1: *Different ABX tasks, where A and X belong to the same category*

| Task | $A$ | $B$ | $X$ |
|---|---|---|---|
| WST | $/bag/_{talker1}$ | $/beg/_{talker1}$ | $/bag/_{talker1}$ |
| AST | $/bag/_{talker1}$ | $/beg/_{talker1}$ | $/bag/_{talker2}$ |

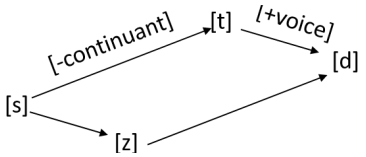


Figure 2: *A hypothetical two-dimensional representation displaying parallelism.*

the phonetic content despite the speaker change, and thus it requires invariance to speaker identity. In Table 1, we illustrate the tasks.

A final score is obtained by averaging ABX scores over all the triplets in the corpus that were tested. In the overall ABX score below, we present the error rate ($1-$ the ABX accuracy score), where an error of zero indicates a representation in which categories are perfectly separated, and an error of $50\%$ represents chance level. In the feature-by-feature ABX score, we use an ABX accuracy score to allow for comparability with the feature-by-feature parallelism analysis.

#### *3.3.2. Parallelism*

The second analysis measures how well the learned representations code individual phonological features, assessing the *parallelism* of the representations [26]. For a given phonological feature, say [voice], representations of the feature are extracted by taking subtractions of phonemes that differ only in this feature (for [voice], [d] $-$ [t], [z] $-$ [s], and so on). In a space with perfect parallelism for voicing, these vectors will be exactly parallel (see Figure 2); the parallelism score we use here assesses relative parallelism, so that all subtraction vectors corresponding to a single feature need only be more parallel (have higher cosines) than pairs of subtraction vectors not corresponding to changes in the same single feature in order to obtain a maximal score (1). They will obtain a minimal score (0) when they are relatively more orthogonal or anti-collinear, and a score of 0.5 when they are no more parallel than vectors corresponding to different features. See [26] for details.

## 4. Experiments

### 4.1. ABX discriminability

Table 2 shows the ABX error rates within and across speaker for audio ($A$), visual ($V$) and the concatenated representation ($A\&V$). The distance used for the ABX tasks is the cosine distance. We notice first that, for all modalities, the ABnet embeddings have a lower error than the raw features. This was already demonstrated for the acoustic modality [6, 27, 28]; we demonstrate that ABnet also improves the visual and concatenated input representations.

Tested on the visual modality, the ABnet improves the across-speaker ABX discriminability. In particular, the multi-task ABnet yields an embedding for visual information that is more discriminative than the (raw) audio features, even though

Table 2: *Within- and across-speaker ABX discrimination (**% error**) for audio, visual and multimodal representations by training condition (raw features, mono-task training, multi-task training).*

| | Raw features | | | Mono-task | | | Multi-task |
|---|---|---|---|---|---|---|---|
| Test | A | V | A&V | A | V | A&V | A&V,A,V,A\|V |
| Within speaker | | | | | | | |
| A | 16.70 | - | - | **8.84** | - | - | 9.80 |
| V | - | 27.54 | - | - | 26.62 | - | **21.78** |
| A&V | - | - | 25.27 | - | - | 10.76 | **10.53** |
| Across speaker | | | | | | | |
| A | 26.12 | - | - | **11.44** | - | - | 12.39 |
| V | - | 40.03 | - | - | 27.12 | - | **24.10** |
| A&V | - | - | 38.33 | - | - | **13.02** | 14.01 |

Table 3: *Pairs of phonemes representing minimal oppositions for phonological features for French.*

| Phonological feature | Pairs |
|---|---|
| Round | [e]-[ø], [ɛ]-[œ], [i]-[y], [ɑ̃]-[ɔ̃] |
| Coronal/Labial | [d]-[b], [s]-[z], [t]-[p], [z]-[v], [n]-[m] |
| Coronal/Dorsal | [d]-[g],[t]-[k] |
| Continuant | [b]-[v],[t]-[s],[d]-[z],[p]-[f] |
| Nasal vowel (NasalV) | [a]-[ɑ̃], [ɛ]-[ɛ̃], [ɔ],[ɔ̃] |
| Nasal consonant (NasalC) | [b]-[m],[d]-[n] |
| Mid/High | [e]-[i],[ø]-[y],[o],[u] |
| Voice | [p]-[b], [t]-[d], [f]-[v], [k]-[g], [s]-[z], [ʃ]-[ʒ] |

the audio signal itself is substantially more discriminative than the visual. Furthermore, for the visual inputs, we reach the best performance with the multi-task model, demonstrating that this model takes advantage of the audio modality present in the training phase to learn a better visual representation.

For the audio modality, the mono-task model achieves the best performance. The presence of the visual modality during training deteriorates the embedding.

To better understand the differences in the ABX scores, we examine the scores for particular phoneme pairs. We look at the pairs that correspond to *minimal oppositions* for phonological features: pairs of sounds that differ only in the given feature. Table 3 lists the oppositions relevant to Hexagonal French. We separate the phonological features into *visual* features—**Round** and **Coronal/Labial**—which correspond to features marked by lip rounding, and thus with clear visual correlates, and the remaining features, which we do not expect to have such strong visual correlates, and which we thus call *non-visual* features.

Figure 3 reports the ABX discriminability (in **% accuracy**) for the pairs in Table 3. We notice that for the non-visual phonological features, the mono-task audio nearly always has the highest ABX accuracy, except in two cases (**Dorsal** and **Continuant**) where the addition of visual information in training (multi-task audio) improves the contrasts slightly. In other words, in this case, representations that include visual information are handicapped with respect to phone discriminability for non-visual features. On the other hand, for the visual phonological features, the audio embeddings have the lowest ABX accuracy. The concatenated embeddings have the highest accu-

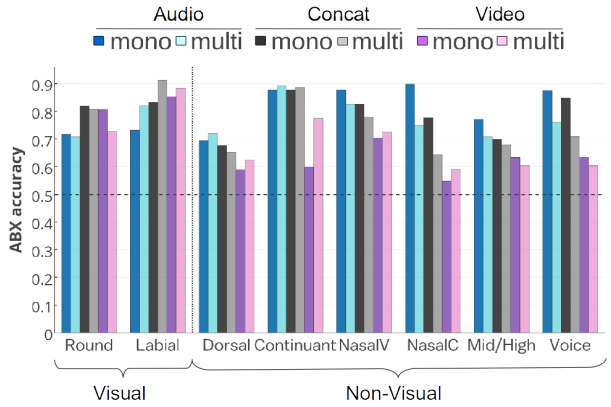


Figure 3: *ABX accuracy, by phonological feature, for different embeddings.*

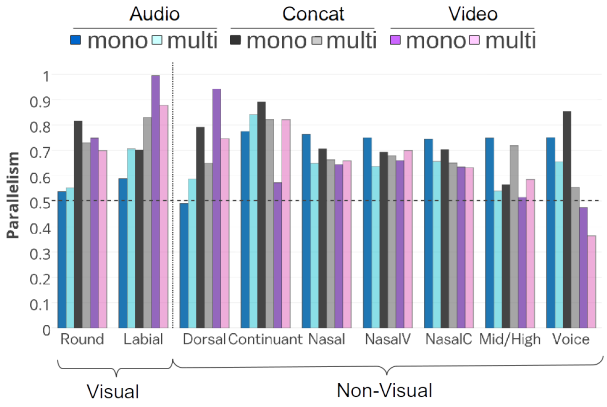


Figure 4: *Parallelism score, by phonological feature, for different embeddings.*

racy. Nevertheless, on average, the multi-task audio embedding gains discriminability compared to mono-task audio embedding for the consonant feature **Coronal/Labial**. Thus, for this feature, the audio inputs benefit from the visual modality present during the training.

### 4.2. Parallelism

We now use the same set of minimal oppositions for the parallelism analysis discussed above, which examines the internal structure of the representations. Figure 4 shows the parallelism scores for the various embeddings. As previously, representations including visual information substantially improve the scores for visual features, indicating that these features have a more consistent representation in these embeddings. For non-visual features, the multi-task audio generally performs better than mono-task audio. However, unlike for the ABX analysis, an acoustic embedding has the best score only for three features (**Nasal vowel, Nasal consonant, Mid/high**). Thus, even for "non-visual" phonological features—those for which no obvious lip movement is expected—representations incorporating visual information have slightly more consistent encodings of these features. On average, the mono-task concatenation embedding has the best parallelism, and is thus the best approximation of a phonological feature representation.

### 4.3. McGurk effect

To verify the plausibility of our models and to test if they can mimic the learning stages of the infant, we can see if they show the same patterns of audio-visual integration as human beings: we test whether presenting an audio signal corresponding to [b] with a mismatched visual signal, corresponding to [g], is perceived by the model as a [d] (the McGurk effect).

We perform an ABX discrimination task to see if these mismatched multi-modal inputs are unexpectedly similar to (hard to distinguish from) [d] (audio and video matched). To assess this, we construct three test sets of multi-modal inputs. The first matches the acoustic [b] with the same, generic visual held-out [b]. The second replaces this with a generic visual held-out [p]. The first gives a baseline score, and the second gives a combination which is mismatched, but which should not show integration effects (the lip movements are the same as for [b]). Finally, the last set is made to simulate the McGurk setting, replacing the visual [b] with a generic held-out [g]. We see that the ABX accuracy scores are lower in the McGurk case (Table 4).

Table 4: *ABX discriminability score between [b] and [d]*

| Task | Visual [b] | Visual [p] | Visual [g] |
|---|---|---|---|
| Mono-task concat | 77.93% | 78.01% | 65.47% |
| Multi-task concat | 93.00% | 92.67% | 79.60% |

## 5. Discussion

This study introduces methods for learning multimodal speech representations in a weakly supervised setting. We use measures of the speech representations' performance on phone discrimination, and introduce analyses of the internal structure of the representations. The discriminability analysis shows that weakly supervised learning using ABnet improves phoneme discriminability over the input features in all cases, and that, for certain phonemic contrasts (those with strong visual correlates), adding visual information helps in discrimination. It also changes the structure of the representation to give more coherent representations of the relevant phonological features. The model can take advantage of the visual information even when it is present only during training. For phonological contrasts which do not benefit from visual information, the methods developed here for adding visual information reduce discriminability in some cases. This shows that visual information should only be exploited selectively, and discarded when it does not provide discriminative information. This suggests future research incorporating into our models a gating or attention system [29] that would learn when to use each modality, or both, in order to dynamically ignore uninformative features at test time.

## 6. Acknowledgment

This work was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Ecole de Neurosciences de Paris, the Region Ile de France DIM Cerveau et pensée, and an Amazon Web Services in Education Research Grant award.

# 7. References


[1] M. J. McGurk H, "Hearing lips and seeing voices," *Nature*, vol. 264, p. 746–748, 1976.

[2] P. K. Kuhl and A. N. Meltzoff, "The intermodal representation of speech in infants," *Infant Behavior and Development*, vol. 7, no. 3, pp. 361–381, jul 1984. [Online]. Available: https://doi.org/10.1016%2Fs0163-6383%2884%2980050-8

[3] L. D. Rosenblum, M. A. Schmuckler, and J. A. Johnson, "The mcgurk effect in infants," *Perception & Psychophysics*, vol. 59, no. 3, pp. 347–357, 1997. [Online]. Available: http://dx.doi.org/10.3758/BF03211902

[4] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 335, no. 1273, pp. 71–78, jan 1992. [Online]. Available: https://doi.org/10.1098%2Frstb.1992.0009

[5] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a"siamese" time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 669–688, 1993. [Online]. Available: http://oro.open.ac.uk/35662/

[6] G. Synnaeve and E. Dupoux, "Weakly supervised multi-embeddings learning of acoustic models," *CoRR*, vol. abs/1412.6645, 2014. [Online]. Available: http://arxiv.org/abs/1412.6645

[7] R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling." in *INTERSPEECH*, 2015, pp. 3179–3183.

[8] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4950–4954.

[9] K. Levin, A. Jansen, and B. Van Durme, "Segmental acoustic indexing for zero resource keyword search," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5828–5832.

[10] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: integrating automatic speech recognition and lip-reading." in *ICSLP*, vol. 94. Citeseer, 1994, pp. 547–550.

[11] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10489-014-0629-7

[12] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 326–338, mar 2016. [Online]. Available: https://doi.org/10.1109%2Ftmm.2016.2520091

[13] J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *CoRR*, vol. abs/1611.05358, 2016. [Online]. Available: http://arxiv.org/abs/1611.05358

[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning."

[15] T. Schatz, "ABX-Discriminability Measures and Applications," Ph.D. dissertation, Ecole Normale Supérieure, Paris, 2016.

[16] J. B. K. L. Galen Andrew, Raman Arora, "Deep canonical correlation analysis," vol. 28, p. 1247–1255, 2013.

[17] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers (IEEE), apr 2015. [Online]. Available: https://doi.org/10.1109%2Ficassp.2015.7178840

[18] M. Versteegh, R. Thiolliere, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015." in *INTERSPEECH*, 2015, pp. 3169–3173.

[19] Y. Benezeth, G. Bachman, G. Le-Jan, N. Souviraà-Labastie, and F. Bimbot, "BL-Database: A French audiovisual database for speech driven lip animation systems," INRIA, Research Report RR-7711, Aug. 2011. [Online]. Available: https://hal.inria.fr/inria-00614761

[20] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, no. 34–47, 2001.

[21] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision (WACV'98)*, ser. WACV '98. Washington, DC, USA: IEEE Computer Society, 1998, pp. 214–. [Online]. Available: http://dl.acm.org/citation.cfm?id=521384.836819

[22] G. Edwards, C. Taylor, and T. Cootes, "Interpreting face images using active appearance models," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. Institute of Electrical and Electronics Engineers (IEEE). [Online]. Available: https://doi.org/10.1109%2Fafgr.1998.670965

[23] H. Sakoe and S. Chiba, "Readings in speech recognition," A. Waibel and K.-F. Lee, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, ch. Dynamic Programming Algorithm Optimization for Spoken Word Recognition, pp. 159–165. [Online]. Available: http://dl.acm.org/citation.cfm?id=108235.108244

[24] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline," in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 1–5.

[25] X.-N. C. F. H. T. Schatz, V. Peddinti and E.Dupoux, "Evaluating speech features with the minimal-pair abx task (ii): Resistance to noise," *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[26] E. Dunbar, G. Synnaeve, and E. Dupoux, "Quantitative methods for comparing featural representations," in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015.

[27] N. Zeghidour, G. Synnaeve, M. Versteegh, and E. Dupoux, "A deep scattering spectrum — deep siamese network pipeline for unsupervised acoustic modeling," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers (IEEE), mar 2016. [Online]. Available: https://doi.org/10.1109%2Ficassp.2016.7472622

[28] N. Zeghidour, G. Synnaeve, N. Usunier, and E. Dupoux, "Joint learning of speaker and phonetic similarities with siamese networks," *Interspeech 2016*, pp. 1295–1299, 2016.

[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473