TOWARDS VISUALLY GROUNDED SUB-WORD SPEECH UNIT DISCOVERY

David Harwath and James Glass

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, Massachusetts, 02139, USA
dharwath@csail.mit.edu, glass@mit.edu

ABSTRACT

In this paper, we investigate the manner in which interpretable sub-word speech units emerge within a convolutional neural network model trained to associate raw speech waveforms with semantically related natural image scenes. We show how diphone boundaries can be superficially extracted from the activation patterns of intermediate layers of the model, suggesting that the model may be leveraging these events for the purpose of word recognition. We present a series of experiments investigating the information encoded by these events.

Index Terms— Vision and language, multimodal speech processing, unsupervised speech processing

1. INTRODUCTION AND PRIOR WORK

The cornerstone of automatic speech recognition (ASR) systems is the taxonomy of discrete, linguistic units modeled by the recognizer. The most salient of these is the vocabulary of words that can be recognized, but "under the hood" many ASR systems utilize a compositional hierarchy of units, e.g. words are composed of phonemes, and phonemes are composed of senone states. This hierarchy of linguistic units is advantageous because it offers flexibility (new words may be specified in the lexicon in terms of existing phonetic models) and data efficiency (phonetic models can be re-used across many different words, allowing for a large degree of parameter sharing between word models). However, it comes at a cost: the training data must be transcribed in terms of the acoustic units, and the compositional mapping between units (e.g. the lexicon mapping phonemes to words) must be specified in advance by an expert linguist. These annotations are expensive to collect, especially for less widely spoken, "low-resource" languages. Unsupervised or weakly-supervised approaches to ASR often attempt to address this problem via the automatic, data-driven discovery of these linguistic units. Some proposed approaches operate at the word level [1, 2, 3], while a separate line of work is concerned with sub-word modeling [4, 5, 6, 7]. Other works have jointly learned sub-word as well as word-level units in a unified framework [8, 9].

A central difficulty faced by unsupervised models of speech is the fact that the acoustic speech waveform is the result of a complex entanglement of many different sources of variability, such as speaker, background noise, reverberation, microphone characteristics, etc. *Self-supervised* models [10] have recently garnered increased attention as an alternative approach to the traditional supervised-unsupervised dichotomy. In lieu of labels, self-supervised learning algorithms leverage informative context found e.g., in another modality. An early example of this is the CELL model introduced by [11] which learned to associate words, represented by phoneme strings, with the visual images they described. Recently, [12, 13, 14] introduced models capable of learning the semantic correspondences between raw speech waveforms and natural images at the pixel level. Subsequent works have continued to explore the leveraging of visual information to guide models of speech audio data [15, 16, 17, 18, 19, 20], and several papers have begun to investigate the nature of the internal representations learned by these visually-grounded models [21, 22]. This paper follows the same general theme, but with a different focus. While [21] and [22] examined the utility of the intermediate representations of visually-grounded speech models to perform tasks such as speaker, phoneme, and word discrimination, they did not investigate if and how discrete, sub-word units may be emerging within the models. Visually-grounded, self-supervised models such as DAVEnet make relatively few assumptions about how sub-word units should be represented. Therefore, if interpretable sub-word unit structure emerges naturally within the network as a by-product of training, the learned structure could provide a fruitful direction for subsequent research on acoustic unit learning. In this paper, we present experiments that suggest that diphone-like structure is being learned by the intermediate layers of the DAVEnet (Deep Audio Visual Embedding network) audio model [20].

2. THE VISUALLY-GROUNDED ACOUSTIC MODEL

For our experiments, we leverage the DAVEnet 5-layer speech CNN described in [20]. This model takes as input log Mel-filterbank spectrograms representing a speech signal, and outputs an embedded representation of the speech intended to capture the high-level semantics of the utterance. This is enforced by training the model to associate

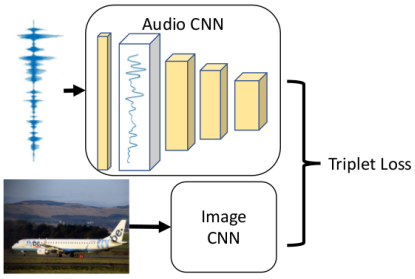

Fig. 1. The DAVEnet audio-visual model. We analyze the activation envelope $e[n]$ of the `conv2` layer.

natural image scenes (encoded with a separate CNN) with spoken captions describing the content of the images (Figure 1). Model performance is measured via Recall@10 on an image/caption retrieval task. The speech model is trained from randomly initialized weights, without any traditional linguistic supervision such as word or phonetic transcriptions, a pronunciation lexicon, etc. As in [20], we train the model on 400,000 image/caption pairs from the Places Audio Caption dataset [20, 13, 23]. We make two small modifications to the model training. First, we employ within-batch semi-hard negative mining [24, 25]. As in [25], we blend the semi-hard negative triplet loss with the standard random-sampled triplet loss; for our experiments we simply weight these terms equally. Second, rather than Matchmap-based similarities, we employ global average pooling to the outputs of both networks, and compute their similarity with a dot product. This is mathematically equivalent to the SISA loss described in [20], but is more computationally efficient for negative mining. Using semi-hard negative mining, we see a boost in Recall@10 from .559 to .641 for image retrieval, and from .506 to .616 for caption retrieval when using a visual model pre-trained on ImageNet [26].

3. EXPERIMENTS

We focus our analysis on the representations learned by the `conv2` layer of DAVEnet, which was previously shown by [21] to encode more phonetic information than other layers of the network. This makes intuitive sense because its receptive field size (125 ms) more closely corresponds to a typical phonetic segment duration (82 ms on the TIMIT database) than the receptive field size of the `conv1` layer (25 ms) or the `conv3` layer (400 ms). When visually examining the outputs of the `conv2` layer, we observed that the activations tend to oscillate with time (Figure 2). This inspired us to investigate the reason for these oscillations in more depth. Given an input spectrogram of N frames, we denote the the activation map output by the second convolutional layer (post-nonlinearity, pre-maxpooling) of DAVEnet as $A \in \mathcal{R}^{Nx256}$, where $A[n, f]$ represents the output of filter f at frame n. We compute the activation envelope signal $e[n]$ by taking the L2 norm across all filter channels, i.e. $e[n] = (\sum_f A[n, f]^2)^{.5}$. Figure 2 depicts $e[n]$ and its associated spectrogram for TIMIT [27] utterance `fisb0_sx49`. An interesting property of $e[n]$ is that it is relatively smooth, and exhibits distinct peaks, indicating that there are particular moments in time that trigger strong activity within the layer. These peaks appear to synchronize with phoneme transitions in the spectrogram, which we validate by applying a simple peak-picking algorithm to the envelope signal and measuring the temporal correspondence between these peaks and the ground-truth phonetic boundary annotations for TIMIT. While any standard peak picking algorithm could be used here, we convolve $e[n]$ with a derivative of Gaussian (DoG) filter (whose shape is controlled via a single hyperparameter σ), i.e. $d[n] = DoG_\sigma[n] * e[n]$. Peaks correspond to positive-to-negative zero crossings in $d[n]$, which are further filtered by a sharpness threshold τ which compares the maximum slope on the rising edge of a peak to the minimum slope on the falling edge; we keep only those peaks for which the difference between these slopes exceeds τ.

Algorithm	Precision	Recall	F1
$e[n]$ peaks	.893	.712	.792
[4]	.764	.762	.763
[29]	.748	.819	.782
[30]	.740	.700	.730

Table 1. Boundary detection on the full TIMIT test. Note that [4] reflects scores on the training set, not the testing set.

Our first experiment measures how well the peaks extracted from $e[n]$ correspond to phonetic boundaries on the full test set of the TIMIT corpus, computing precision, recall, and F1 against the ground-truth boundaries. We use the Places audio caption DAVEnet model as-is, and do not do any further training or adaptation on the TIMIT data. We follow [28] and use a 20ms tolerance window for boundary detection. We performed a grid search over τ and σ, and achieved a maximum F1 of .792 at $\tau = 0.15$ and $\sigma = 0.5$; but performance was not very sensitive to these exact settings. In Table 1, we compare against several published approaches for blind phone boundary detection. Our method outperforms all of them in terms of F1 score, but does not constitute a fair comparison because our model underwent self-supervised training on the Places audio captions. The key takeaway is the fact that $e[n]$ performs very well as a phone boundary detector despite never being explicitly trained to do so.

Given that some regions of an input spectrogram give rise to peaks in the activation pattern of the `conv2` layer of DAVEnet, our second experiment examines to what extent the DAVEnet model is leveraging the information contained in these peaks to perform cross-modal retrieval. Here, we fix all of the weights of the DAVEnet model *except* the bias vector of the `conv3` layer. We then insert an ablation layer between `conv2` and `conv3`. This layer computes the peaks in the $e[n]$ signal and then uses them to create a mask matrix

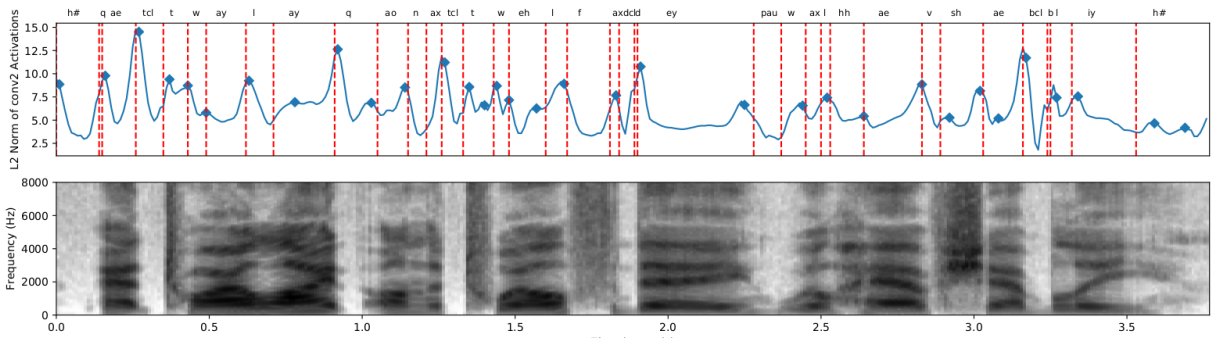

Fig. 2. The spectrogram for TIMIT utterance `fisb0_sx49` (bottom) and its associated $e[n]$ signal (top, blue curve). Peaks found in $e[n]$ shown by blue diamonds, and ground-truth TIMIT phone boundaries denoted by the vertical dashed red lines).

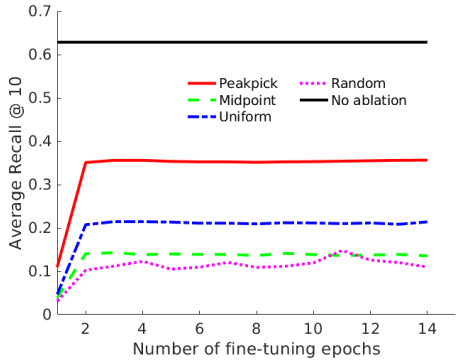

Fig. 3. R@10 scores using various ablation methods.

$M \in \mathcal{R}^{Nx256}$, where $M[n, f] = 1$ if $e[n]$ has a peak at n, and $M[n, f] = 0$ otherwise. The ablated outputs of `conv2` are then computed as $\hat{A} = M \odot A$, and $\hat{A}$ is fed as input to the subsequent layers of the network. Because the ablation layer changes the magnitude of the summed input seen by each neuron in the `conv3` layer, we fine-tune only the bias of this layer on the ablated outputs from `conv2`, using the image and caption ranking objective described in Section 2. By comparing the retrieval recall scores achieved by the ablated network against those of the original model, we can infer to what extent the convolutional filters of the already-trained DAVEnet model have learned to focus on the `conv2` activation peaks, as opposed to other regions of the `conv2` output. Figure 3 displays the average of the image-to-caption and caption-to-image recall @ 10 scores as a function of the number of fine-tuning epochs. The horizontal black line represents the recall score when no ablation is used (0.629), and the red solid line shows the score achieved when using the peak-picking based ablation. After a single epoch of fine-tuning, the R@10 score rebounds from 0.117 to 0.351, where it remains constant. We compare the peak-picking ablation against uniform sampling, random sampling, and sampling the midpoint frame between each consecutive pair of $e[n]$ peaks. For uniform and random sampling, we keep the number of ablated frames constant for each utterance; if a given utterance of length N was found to have N_p peaks in its $e[n]$ signal, then we retain N_p uniformly-spaced or randomly sampled `conv2` activation frames across the utterance. On average, 1 out of every 11.84 frames was found to be a peak, meaning that 91.6% of the `conv2` output frames were set to zero. In Figure 3, we see that while all ablation methods suffer a loss in retrieval accuracy, the peak-picking ablation model still achieves 60% of the performance of the non-ablated model. All other methods fare worse, indicating that the filters of the DAVEnet audio model have learned to leverage the $e[n]$ peaks for word discrimination much more than other parts of the speech signal.

Thus far, we have shown that the `conv2` layer of the DAVEnet audio model is highly sensitive to specific regions of an input spectrogram, that these regions are especially informative for inferring the semantics (and thus the lexical content) of an utterance, and that these regions tend to occur at the transition point between two phones. Our last experiments investigate the geometry of the embedding space in which these activation peaks reside. We first extracted a total of 39,871 peaks for the 1,344 utterances comprising the TIMIT complete test set. We represent each peak with its corresponding 256-dimensional embedding vector produced by the `conv2` layer of DAVEnet. We then assign a label to each peak according to the ground-truth sequence of phones that fall within a 40 ms window around the peak. We follow a similar scheme to the 39-phone mapping for TIMIT, but map stop closures to their associated stop phoneme instead of silence. Under this mapping, we found that approximately 18.1% of the peaks fell within a single phone segment, 76.5% of peaks captured a diphone boundary, and 5.3% of the peaks overlapped three phones. In addition to a phonetic label for each peak, we also derive a manner label by mapping each phone to its associated broad manner class (vowel, stop, nasal, fricative, semivowel, affricate, flap, and silence). We projected the peak embeddings down to 2 dimensions using PCA, and plot the peaks corresponding to the 10 most frequently occurring broad manner class labels in Figure 4. We can see that

peaks belonging to the same underlying manner class cluster together quite well. Furthermore, we observe that CV, V, and VC syllable structure is captured by the first principal component; vowel_stop, vowel_fricative, and vowel_nasal peaks are concentrated on right-hand side of the space, while stop_vowel, fricative_vowel, and nasal_vowel reside on the left-hand side of the space.

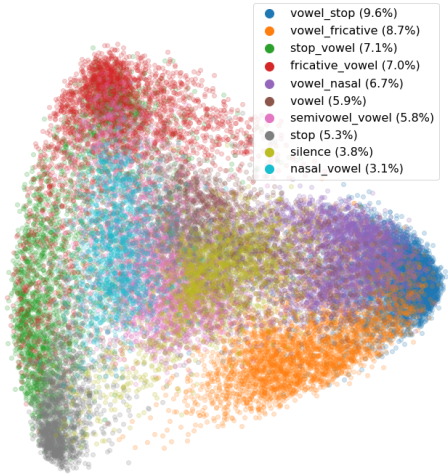

Fig. 4. PCA analysis of $e[n]$ peaks extracted from the TIMIT full test set.

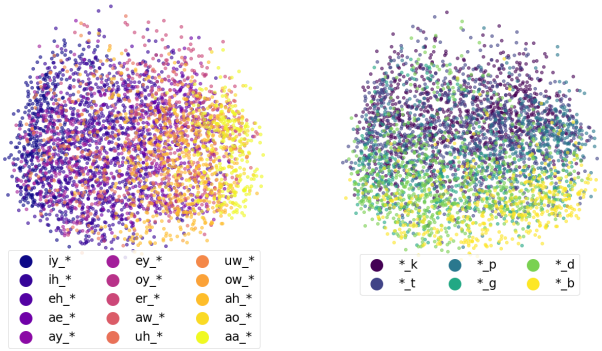

Fig. 5. PCA analysis of the peaks corresponding to vowel-stop transitions.

In Figure 5, we examine the properties of peaks belonging to the vowel_stop class in more detail. We compute a second PCA transform specific for this class, and plot the associated peaks along their first two principal components. In the left-hand scatter plot, the peaks are color coded according to the vowel in the left context of the peak; in the right-hand plot, they are color coded according to their right-hand stop context. Broadly speaking, the first principal component seems to select for frontness of the vowel, while the second component captures the voicing of the stop consonant.

Our qualitative analysis suggests that the discovered peaks cluster by manner class at a coarse scale, and by phonetic identity at a finer scale. We quantify this by clustering the peak vectors using K-means, and computing the adjusted mutual information (AMI) [31] between the clustering output and the underlying phone and manner class sequences for each peak (Figure 6). AMI is maximized for the manner class label sequences at $K = 40$ clusters and steadily falls off as more clusters are specified, while the AMI for phonetic labels plateaus between 200 and 700 clusters.

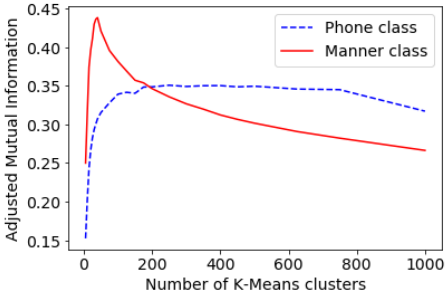

Fig. 6. Adjusted mutual information between K-means clustering of peaks and their underlying phone and manner class sequences.

4. CONCLUDING DISCUSSION

In this paper, we investigated the encoding of sub-word information in the `conv2` layer of the DAVEnet visually-grounded speech model. We observed that the magnitude of the activations within this layer tend to exhibit local maxima at diphone transitions, which we quantified by using these maxima to detect phone boundaries on TIMIT. Furthermore, we performed ablation experiments for an image/caption retrieval that suggested that the DAVEnet audio model leverages the $e[n]$/"diphone" peaks more than other regions of the signal for the purpose of word recognition. Finally, we examined the geometry of the space occupied by the peak embedding vectors and found the emergence of clusters of diphone units which share broad phonetic manner class membership; within these clusters, different dimensions appear to correlate with distinctive features such as vowel frontness or stop voicing. In our future work, we plan to further explore the topic of leveraging visually-grounded acoustic models to discover discrete, pseudo-linguistic units. We would like to explicitly incorporate mechanisms into DAVEnet for inferring a discrete, compositional hierarchy of interpretable phone-like, syllable-like, word-like, phrase-like, etc. units that could provide a rich account of a spoken language in a self-supervised fashion. Finally, we believe that our ablation analysis in Section 3 points the way towards a non-linear downsampling scheme that would enable acoustic observation sequences to be more closely aligned with phone or character sequences, which could find application in supervised ASR systems.

5. REFERENCES

[1] Alex Park and James Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[2] Aren Jansen, Kenneth Church, and Hynek Hermansky, "Toward spoken term discovery at scale with zero resources," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2010.

[3] Herman Kamper, Aren Jansen, and Sharon Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *Computer Speech and Language*, vol. 46, no. 3, pp. 154–174, 2017.

[4] Chia-Ying Lee and James Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012.

[5] Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008.

[6] Lucas Ondel, Lukaš Burget, and Jan Černocký, "Variational inference for acoustic unit discovery," 2016.

[7] Aren Jansen, Samuel Thomas, and Hynek Hermansky, "Weak top-down constraints for unsupervised acoustic model training," in *ICASSP*, 2013.

[8] Herb Gish, Man-Hung Siu, Arthur Chan, and William Belfield, "Unsupervised training of an HMM-based speech recognizer for topic classification," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2009.

[9] Chia-ying Lee, Timothy O'Donnell, and James Glass, "Unsupervised lexicon discovery from acoustic input," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 389–403, 2015.

[10] Virginia de Sa, "Learning classification with unlabeled data," 1994.

[11] Deb Roy and Alex Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science*, vol. 26, pp. 113–146, 2002.

[12] David Harwath and James Glass, "Deep multimodal semantic embeddings for speech and images," in *Proc. IEEE Workshop on Automfatic Speech Recognition and Understanding (ASRU)*, 2015.

[13] David Harwath, Antonio Torralba, and James R. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. Neural Information Processing Systems (NIPS)*, 2016.

[14] Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux, "Learning words from images and speech," in *In NIPS Workshop on Learning Semantics*, 2014.

[15] David Harwath and James Glass, "Learning word-like units from joint audio-visual analysis," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[16] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," in *INTERSPEECH*, 2017.

[17] Grzegorz Chrupala, Lieke Gelderloos, and Afra Alishahi, "Representations of language in a model of visually grounded speech signal," in *ACL*, 2017.

[18] Odette Scharenborg, Laurent Besacier, Alan W. Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux, "Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "speaking rosetta" JSALT 2017 workshop," *CoRR*, vol. abs/1802.05092, 2018.

[19] David Harwath, Galen Chuang, and James Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *ICASSP*, 2018.

[20] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass, "Jointly discovering visual objects and spoken words from raw sensory input," 2018.

[21] Jennifer Drexler and James Glass, "Analysis of audio-visual features for unsupervised speech recognition," in *Grounded Language Understanding Workshop*, 2017.

[22] Afra Alishahi, Marie Barking, and Grzegorz Chrupala, "Encoding of phonology in a recurrent neural model of grounded speech," in *CoNLL*, 2017.

[23] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *Proc. Neural Information Processing Systems (NIPS)*, 2014.

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," 2018.

[25] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R Channing Moore, and Rif A Saurous, "Unsupervised learning of semantic audio representations," 2018.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[27] John Garofolo, Lori Lamel, William Fisher, Jonathan Fiscus, David Pallet, Nancy Dahlgren, and Victor Zue, "The TIMIT acoustic-phonetic continuous speech corpus," 1993.

[28] Odette Scharenborg, Vincent Wan, and Mirjam Ernestus, "Unsupervised speech segmentation: An analy- sis of the hypothesized phone boundaries," *The journal of the acoustical society of america (JASA)*, vol. 127, pp. 1084–1095, 2010.

[29] Paul Michel, Okko Räsänen, Roland Thiolliére, and Emmanuel Dupoux, "Blind phoneme segmentation with temporal prediction errors," in *Proceedings of the ACL Student Research Workshop*, 2017.

[30] Okko Räsänen, "Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level.," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2014.

[31] Nguyen Xuan Vinh, Julien Epps, and James Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Dec. 2010.