Multimodal Word Discovery and Retrieval with Spoken Descriptions and Visual Concepts

Liming Wang, *Student Member, IEEE,* Mark Hasegawa-Johnson, *Member, IEEE,*

***Abstract*—In the absence of dictionaries, translators, or grammars, it is still possible to learn some of the words of a new language by listening to spoken descriptions of images. If several images, each containing a particular visually salient object, each co-occur with a particular sequence of speech sounds, we can infer that those speech sounds are a word whose definition is the visible object. A multimodal word discovery system accepts, as input, a database of spoken descriptions of images (or a set of corresponding phone transcriptions) and learns a mapping from waveform segments (or phone strings) to their associated image concepts. In this paper, four multimodal word discovery systems are demonstrated: three models based on statistical machine translation (SMT) and one based on neural machine translation (NMT). The systems are trained with phonetic transcriptions, MFCC and multilingual bottleneck features (MBN). On the phone-level, the SMT outperforms the NMT model, achieving a 61.6% F1 score in the phone-level word discovery task on Flickr30k. On the audio-level, we compared our models with the existing ES-KMeans algorithm for word discovery and present some of the challenges in multimodal spoken word discovery.**

***Index Terms*—unsupervised word discovery, language acquisition, machine translation, multimodal learning**

I. Introduction

Unsupervised word discovery aims to segment and cluster spoken sentences, or their corresponding phone transcriptions, into words. Word discovery is a useful first step in the development of speech technology for unwritten languages or languages in which word segmentation and lexicon will be prohibitively expensive to acquire. Unsupervised word discovery system exists [1]–[3], but the task has been shown to be quite challenging. In this paper, we explore the use of image as an additional information source for word discovery: if each utterance is known to be a spoken description of an image, the image labels can be seen as a bag of noisy word labels for the speech. As a result, we can discover word units from unsegmented spoken sounds by utilizing the co-occurrence patterns between sound units and the image labels, as illustrated in Fig. (1).

II. Related works

Several works have used raw audio to discover word units. Methods that imitate child language acquisition often begin by finding recurring patterns in audio [2], [4]. Non-parametric Bayesian hidden Markov models (HMMs) have been widely used in word-unit discovery and various other clustering problem with audio, e.g., a latent Dirichlet process with HMM acoustic models can be used to jointly segment and cluster raw audio into sub-word units [5], [6], or the HMM can be regularized using an ℓ_p norm as sparsity constraint to encourage purer clusters [1]. Using word embeddings as features, it is possible to perform automatic word discovery by modeling each word as a Gaussian mixture model (GMM) with a Dirichlet prior on its parameters; the model can be trained using expectation maximization (EM) with Gibbs sampling, or using a weighted K-Means algorithm [3]. The segmental embedded systems in [3] outperformed the previous systems by 10 % boundary F-score and 30 % word token F-score for the low-resource language Xitsonga during the 2017 zero-resource speech challenge (ZRSC) [7]. Other works have focused on discovering word units from phone sequences or character sequences, such as models based on Pitman-Yor processes [8].

A related task to the unsupervised spoken word discovery is query-by-example keyword search in audio, which aims to only search for a collection of keywords and leaves the rest of the speech as background. The most recent widely published benchmark evaluation of this task was the NIST OpenKWS evaluation set on the language Georgian. The Kaldi OpenKWS system [9] trained a Deep Neural Network (DNN)-HMM hybrid system and decoded the out-of-vocabulary (OOV) queries by fusing the decoding scores on the word-level, phonetic-level and morpheme-level lattices to maximize the average-term weighted value. The BBN system [10] combined several acoustic models based on DNN, long short-term memory (LSTM) and convolutional neural net (CNN) on subword units to perform joint decoding and handled OOV queries on the sub-word unit. The STC keyword search system [11] combined 9 different acoustic models based on DNN and GMM with a phone-posterior based OOV decoder [12].

A multilingual approach for spoken word discovery has been proposed by [13], which developed a variant of the IBM model 3 SMT to discover word units of an under-resourced language by aligning parallel texts in a high-resourced language. The same task has been attempted [14] using NMT with attention [15] to align speech or phone sequences to the word labels of the high-resourced language; modifications of the attention mechanism ensure coverage and richer context. If the true phone sequence in the under-resourced language is unknown, pseudo-phone labels generated by an unsupervised non-parametric Bayesian model [6] can be used as input to the NMT [16].

The database used in this paper was first published as

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

Manuscript received October 1, 2019; revised Februrary 3, 2020 and May 14, 2020; accepted May 17, 2020. The associate editor

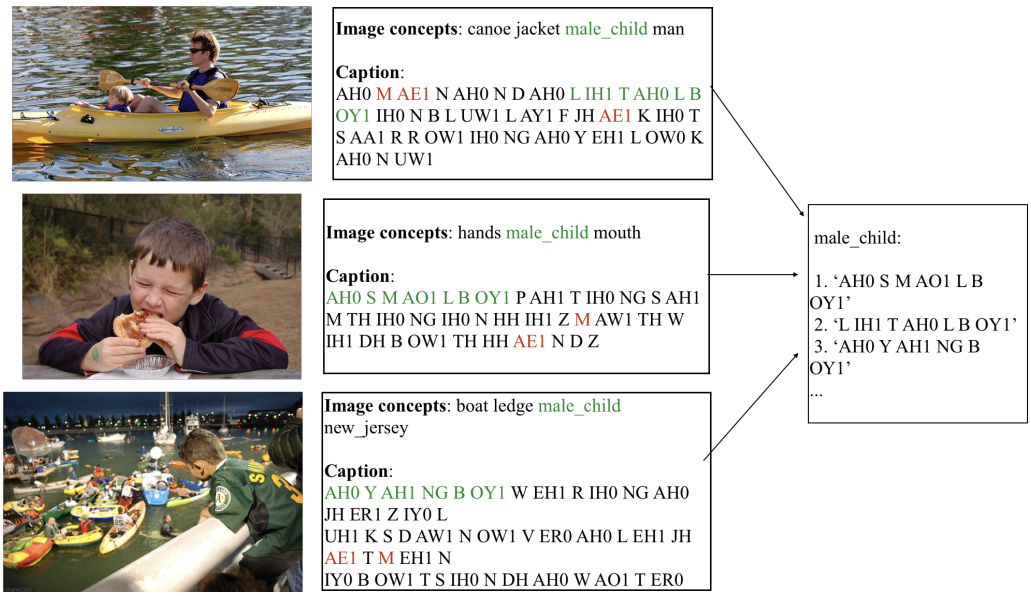

Fig. 1. Example of multimodal word discovery: The concept "male child" is learned by finding co-occurrences between phone strings and image labels. The phones in green of the caption represents the ground truth phrases for the image concept/label of "male child" (marked in green as well), while the phones in red represents the potential confounder `M AE1 N` to the algorithm. Notice that no word boundary or exact word label is provided and there can be multiple image labels per image. Word discovery algorithms that require sequential phones would find the confounding sequence `M AE1 N` in only the first image, while those that permit non-sequential phones would find it in all three images.

an image captioning corpus, for which the baseline system [17] used IBM model I and II [18] combined with Kernel Canonical Component Analysis (KCCA) for mapping both image and text to a joint space. [19], [20] developed a two-branch neural network system to learn the joint representation of image and text. The speech files were first used to train an end-to-end image retrieval system from speech [21]–[24]: Pretrained image embeddings provided supervision to learn the acoustic embeddings of spoken captions by maximizing cosine similarity between the embeddings of image-caption pairs. The task of multimodal word discovery was, we believe, first proposed in [25], where it was performed as a generalization of the image retrieval problem: [25] found that the same acoustic embedding used in the image retrieval systems can be used to discover word-like segments in speech and bounding boxes in an image by exhaustively searching over grids in the speech and image. The exhaustive search was replaced in [26] by a more efficient convolutional time alignment, in which peaks in the similarity between the image and audio convolutional networks were taken to indicate discovered image concepts and audio words, respectively. Convolutional multimodal time alignment is able to automatically discover word alignments between Hindi and English [27], and to discover phone-like units in speech [28]. In a previous paper [29], we have reformulated the multimodal word discovery problem as a statistical machine translation problem to align between a sequence of phone labels and a set of image concepts matched one-to-one to the image regions. In [29], however, the phone labels are assumed to be discrete and we only explored the alignment model with the mixture assumption. This paper extends the previous paper to allow any unsegmented, continuous acoustic features for the audio modality.

III. Notations

In the paper, we use $N(x|\mu, \Sigma)$ to denote the multivariate normal distribution with mean μ and covariance matrix Σ evaluated at point x. When there is a random process $(U_1, X_1) \cdots (U_T, X_T)$, $c_t(u; x)$ denotes the count at the t-th random variable pair (U_t, X_t). Both probability mass functions and probability density functions are written with the notation $p(x)$, and thus $p(x)$ may represent either $\mathbf{Pr}\{X = x\}$ or $\frac{d}{dx}\mathbf{Pr}\{X \leq x\}$, depending on whether X is discrete or continuous. The parameters of the distribution are often omitted for simplicity: $p(\cdot; \theta) =: p(\cdot)$. Further, i.i.d stands for "independent, identically distributed" in the subsequent sections.

$c(u; x)$ is the *count* of occurrences in which random variable U takes value u when X is observed to be x. $c(u|v; x)$ represents the number of times random variable U with value u is *aligned* to random variable V with value v, which is equivalent to $c(i; x)$ where I is the alignment variable between U and V. $\mathbb{E}_x[\cdot]$ denotes the expectation over variable X and $\langle\cdot\rangle$ is a shorthand to denote the expectation of the counts. For vectors and sequences, we use $x_{s:t}$ to represent the elements from index s to t.

IV. Problem formulation

A. Multimodal word discovery

Suppose we have a sequence of acoustic feature frames or phone symbols $x_1, \ldots, x_{T_x}$, $x_t \in \mathcal{X}$ and a set of image concepts $y_0, y_1, \ldots, y_{T_y}$, $y_i \in \mathcal{Y} \cup \{\texttt{NULL}\}$ with $y_0 \equiv \texttt{NULL}$. The goal of our algorithm is a sequence learning problem to align every sequence of audio feature frames $\mathbf{x} = [x_1, \ldots, x_{T_x}]$ to a sequence of image concept $\mathbf{y} = [y_1, \ldots, y_{T_y}]$. We assume that $T_x > T_y$. In other words, we seek an alignment matrix

$A \in [0,1]^{T_x \times T_y} = [\mathbf{a}_1^\top, \ldots, \mathbf{a}_{T_x}^\top]^\top = [\tilde{\mathbf{a}}_1 \ldots \tilde{\mathbf{a}}_{T_y}]$, such that the following *many-to-one* constraint are met:

$$a_{ti} \geq 0, \ \sum_{i=0}^{T_y} a_{ti} = 1, \ t = 1, \ldots, T_x, \tag{1}$$

which intuitively ensures that one feature frame is aligned to one and only one concept. For audio frames that are unaligned to any of the image concepts, the `NULL` symbol acts as a placeholder concept to enforce the constraint. Further, SMT-based models assume that $a_{ti} \in \{0, 1\}$, in which case we define $i(t)$ to be the index of the image concept aligned to the frame at time t, i.e., $a_{ti(t)} = 1$, and let $\mathcal{A}$ be the set of all matrices that satisfy Eq. (1). A maximum-likelihood alignment model tries to maximize:

$$p(\mathbf{x}, \mathbf{y}|\mathbf{\Theta}) = \sum_{\mathbf{A} \in \mathcal{A}} p(\mathbf{A}|\Theta) p(\mathbf{x}, \mathbf{y}|\mathbf{A}, \Theta), \tag{2}$$

over Θ for each input-concept sequence pair. When $\mathcal{X}$ is a finite set of ground truth phonetic symbols for the spoken language, the problem is described later as *phone-level* word discovery; when $\mathcal{X} = \mathbb{R}^D$, where D is the dimension of some acoustic features such as mel-frequency ceptral coefficients [30], bottleneck features [31], [32] or acoustic embeddings, the problem is described as *audio-level* word discovery. For the audio-level word discovery, we largely ignore speaker variation and assume that proper speaker adaptation techniques have been applied before the processing of our models.

B. Statistical multimodal alignment models

The statistical multimodal alignment models (SMA) learn to generate phone sequences from image concepts and assume that $p(\mathbf{y}|\Theta)$ does not depend on Θ. The assumption of a model-independent class distribution does not limit the performance since the decision of the model is based on the posterior alignment probabilities:

$$p(\mathbf{A}|\mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{A}) p(\mathbf{x}, \mathbf{y}|\mathbf{A})}{\sum_{\mathbf{A}' \in \mathcal{A}} p(\mathbf{A}') p(\mathbf{x}, \mathbf{y}|\mathbf{A}')} \tag{3}$$

$$= \frac{p(\mathbf{A}) p(\mathbf{x}|\mathbf{y}, \mathbf{A})}{\sum_{\mathbf{A}' \in \mathcal{A}} p(\mathbf{A}') p(\mathbf{x}|\mathbf{y}, \mathbf{A}')}, \tag{4}$$

which does not involve $p(\mathbf{y})$. Further, such assumption makes maximizing Eq. (2) equivalent to maximizing the likelihood $p(\mathbf{x}|\mathbf{y})$:

$$\begin{aligned} \arg\max_{\Theta} p(\mathbf{x}, \mathbf{y}|\mathbf{\Theta}) &= \arg\max_{\Theta} p(\mathbf{x}|\mathbf{y}, \mathbf{\Theta}) \\ &= \arg\max_{\Theta} \sum_{\mathbf{A} \in \mathcal{A}} p(\mathbf{A}|\Theta) p(\mathbf{x}|\mathbf{y}, \mathbf{A}, \Theta). \end{aligned} \tag{5}$$

C. Word discovery with neural multimodal alignment models

Instead of having a model-independent distribution over the image concepts, the neural multimodal alignment model (NMA) assumes a uniform distribution over the acoustic feature frames:

$$p(\mathbf{x}|T_x) = \frac{1}{|\mathcal{X}|^{T_x}}. \tag{6}$$

Therefore, maximizing Eq. (2) is equivalent to maximizing the posterior probability:

$$\arg\max_{\mathbf{\Theta}} p(\mathbf{y}|\mathbf{x}, \Theta) = \arg\max \int_{\mathbf{A} \in \mathcal{A}} p(\mathbf{A}|\mathbf{x}, \Theta) p(\mathbf{y}|\mathbf{x}, \mathbf{A}, \Theta) d\mathbf{A}, \tag{7}$$

where $\mathcal{A}$ is the set of real-valued matrices that satisfies Eq. (1). We have proposed several SMA and NMA models for both phone-level and audio-level discovery and their relations are shown in Fig. (2).

D. Multimodal image retrieval

In a standard multimodal image retrieval problem [19], we have a database with image-caption pairs $(\mathbf{x}^1, \mathbf{y}^{\pi(1)}), \ldots, (\mathbf{x}^S, \mathbf{y}^{\pi(S)})$, where image $\mathbf{y}^{\pi(i)}$ is uniquely described by sentence $\mathbf{x}^i$. The goal is to learn the one-to-one mapping $\pi : \{1, \ldots, S\} \rightarrow \{1, \ldots, S\}$ that maps from each caption to the corresponding image.

V. Simplified mixture multimodal alignment models

At the first glance, Eq. (5) may seem daunting to learn since it contains the summation of $T_x^{T_y+1}$ terms and can be hard to break down into a reasonable number of parameters. However, if each feature frame is assumed to be independently generated by a single concept selected uniformly from the set of image concepts, the expression can be broken nicely into the translation probabilities between single pairs of concepts and acoustic frames/phonetic labels. Such a model is referred later as a *simplified mixture model*.

A. Phone-level model

For the phone-level word discovery, following [33], we can make use of the following assumptions:

1) *Hard alignment assumption*: $\mathbf{A}$ is integer-valued, specifically $a_{ti} = 1$ for $i = i(t)$, else $a_{ti} = 0$;
2) *Uniform prior assumption*: all alignments are equally likely given only $\mathbf{y}$: $p(\mathbf{A}|\mathbf{y}, T_x) = \frac{1}{(T_y+1)^{T_x}}$;
3) *Mixture assumption*: given the alignment, each phone depends only on its aligned image concept, thus $p(x_t|x_{1:(t-1)}, \mathbf{A}, \mathbf{y}) = p(x_t|y_{i(t)})$. A consequence of this assumption is that we lose the ability to model phone sequence information; each word is modeled as a mixture of phones, with no ability to model the sequence information that distinguishes any word from its own temporal permutations.

Eq. (5) is then simplified to:

$$\frac{1}{(T_y+1)^{T_x}} \prod_{t=1}^{T_x} \sum_{i(t)=0}^{T_y} p(x_t|y_{i(t)}). \tag{8}$$

SMT model using similar assumptions are called *mixture model* mainly due to the independence between alignments. Optimization with EM results in an iterative formula in terms

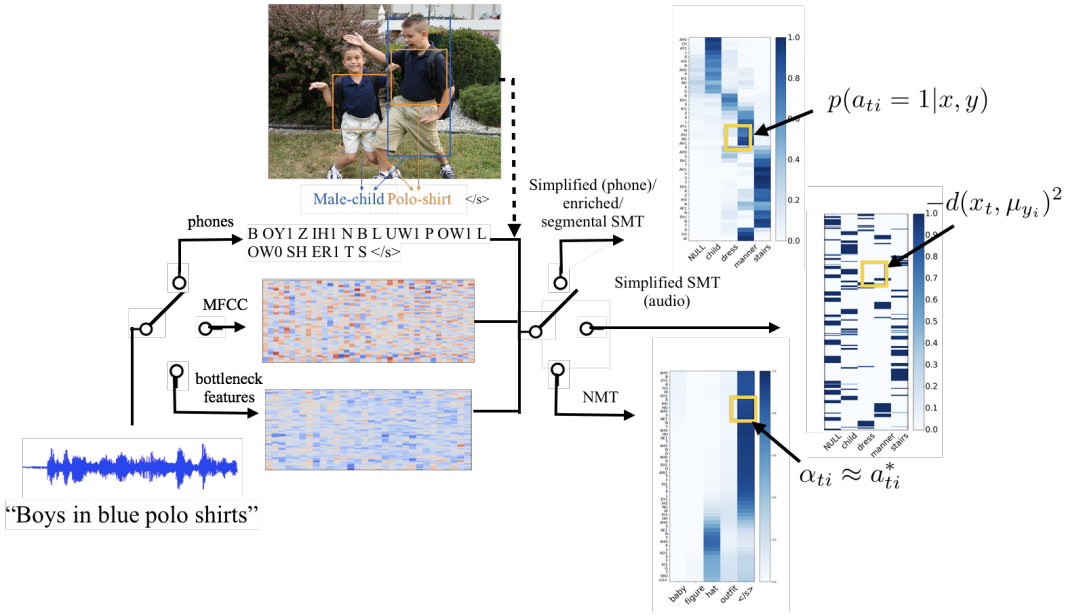

Fig. 2. Comparison of various multimodal word discovery systems. The input of the models are either the ground truth phonetic transcription, the MFCC feature and the bottleneck feature of the spoken caption, while the output is the predicted alignment matrices, which is the negative sum square distance for the simplified model, the alignment posterior probability for the enriched model and the attention weights for the NMT.

of the expected counts of a given phone-concept alignment [18]:

$$\frac{\langle c(x_t|y_i;\mathbf{x},\mathbf{y})\rangle}{\sum_{i'=1}^{T_y}\langle c(x_t|y_{i'};\mathbf{x},\mathbf{y})\rangle} = \frac{p(x_t|y_i)}{\sum_{i'=1}^{T_y} p(x_t|y_{i'})}. \tag{9}$$

The optimal alignment between the phones and image concepts can be then obtained by finding the highest-scored translation pair of a given sentence:

$$i^*(t) = \arg\max_i p(x_t|y_i). \tag{10}$$

B. Audio-level model

Continuous acoustic features may be modeled as a cluster-mixture model, in which each image concept y_i is modeled as a set of acoustic feature vectors with centroids $\{\mu_m(y_i)\}_{i=1,m=1}^{T_y,M}$. Distinct cluster centroids may model segmental variation (multiple phones are segmented to form a word) and/or production variation (any given phone may be pronounced in several different ways). In the cluster-mixture model, both segmental variation and production variation are modeled using the same mechanism. Instead of maximizing the likelihood function, the model tries to minimize:

$$\min_{\mathbf{A}\in\mathcal{A},\mathbf{m}\in\mathcal{M}^{T_x}} \sum_{t=1}^{T_x} \|x_t - \mu_{m(t)}(y_{i(t)})\|_2^2, \tag{11}$$

where $\mathcal{M} = \{1,\cdots,M\}$. The update is similar to the standard K-Means algorithm but guided by the image concepts corresponding to each utterances:

$$i^*(t), m^*(t) = \arg\min_{i,m} \|x_t - \mu_m(y_i)\|^2 \tag{12}$$

$$\mu_m(y_i) = \frac{\sum_{t:i(t)=i,m(t)=m} x_t}{c(i,m;\mathbf{x},\mathbf{y})}. \tag{13}$$

C. Time-dependent alignment probabilities

Bock and her colleagues [34] have demonstrated that the order of eye fixations on an image predicts the order in which the objects will be mentioned. If we speculate that readers of a left-to-right orthography (such as English) tend to also read images left-to-right, it follows that native speakers of English may show a tendency to generate sentences that name the objects in an image from left to right. Brown et. al. [33] has proposed an extension to the simplified mixture model capable of discovering a left-to-right description bias, if it exists: they proposed to relax the uniform prior assumption by making the alignment probability time-dependent. Specifically they introduce a set of parameters $p(i|t,T_x,T_y) := \mathbf{Pr}\{i(t) = i|\mathbf{x}_{1:t},\mathbf{y}_{1:i},T_x,T_y\}$, $i \in \{1,\ldots,T_y\}$, $t \in \{1,\ldots,T_x\}$ with the constraint that $\sum_{i=0}^{T_y} p(i|t,T_x,T_y) = 1$. As a result, Eq. (5) is modified to be:

$$\frac{\epsilon}{(T_y+1)^{T_x}} \prod_{t=1}^{T_x} \sum_{i(t)=0}^{T_y} p(i(t)|t,T_x,T_y) p(x_t|y_{i(t)}), \tag{14}$$

The expected count is then a weighted version of the simplified mixture counterpart:

$$\frac{\langle c(x_t|y_i;\mathbf{x},\mathbf{y})\rangle}{\sum_{i'=1}^{T_y}\langle c(x_t|y_{i'};\mathbf{x},\mathbf{y})\rangle} = \frac{p(i|t,T_x,T_y)p(x_t|y_i)}{\sum_{i'=1}^{T_y} p(i'|t,T_x,T_y)p(x_t|y_{i'})}. \tag{15}$$

D. Gaussian mixture models

The cluster-mixture alignment model (Eq. (11)) is reasonable if this difference between a feature vector and its cluster centroid is a Gaussian random variable, with spherical covariance whose radius is independent of the cluster index.

A slightly richer model is a Gaussian mixture model (GMM), used to model each image concept:

$$p(x_t|y_i) = \sum_{m=1}^{M} c_m(y_i) N(x_t|\mu_m(y_j), \Sigma_m(y_j)), \tag{16}$$

where $\{c_m(y)\}_{i=1}^{M}$ is the prior distribution for each mixture associated with concept y_i. Under the same assumption (1)-(3) as the discrete mixture model, the expected count of the continuous mixture model takes the form:

$$\frac{\langle c_t(i,m;\mathbf{x},\mathbf{y})\rangle}{\sum_{i',m'}\langle c_t(i',m';\mathbf{x},\mathbf{y})\rangle} = \frac{c_m(y_i)N(\mathbf{x}_t|\mu_m(y_i),\Sigma_m(y_i))}{\sum_{i'=1}^{T_y}\sum_{m=1}^{M} c_m(y_{i'})N(\mathbf{x}_t|\mu_m(y_{i'}),\Sigma_m(y_{i'}))} \tag{17}$$

$$\langle c_t(x_t|y_i;\mathbf{x},\mathbf{y})\rangle = \sum_{m=1}^{M}\langle c_t(i,m;\mathbf{x},\mathbf{y})\rangle. \tag{18}$$

The EM algorithm for a GMM tries to maximize the *Baum auxiliary function*:

$$Q(\bar{\Theta},\Theta) = \sum_{t=1}^{T_x}\sum_{i(t)=0}^{T_y}\sum_{m(t)=1}^{M} p(i(t),m(t)|\mathbf{x},\mathbf{y},\bar{\Theta})\cdot \log p(x_t,i(t),m(t)|\mathbf{y},\Theta) \tag{19}$$

Although not the logarithm of the translation probability as in the discrete case, this auxiliary function is known to provide a lower bound for the log-translation probability and guarantees to converge to a local optimum [35]. The main advantage of using the Baum auxiliary objective is that the mixture means and variances now have closed-form expressions in term of the expected count and the acoustic feature frames. The auxiliary function approach will be applicable to all the audio-level SMA models with continuous features in the subsequent sections. More details about the EM update with Baum auxiliary function can be found, for example, in [35].

In fact, the K-Means-based algorithm in the previous section can be seen as a simplification of the enriched mixture model when the mixture distribution is a Gaussian mixture with diagonal and asymptotically zero covariances for each component.

VI. Segmental multimodal alignment models

One major bottleneck of performance for a mixture model is its independence assumption. The alignments for each feature frame are assumed to be independent and thus the global patterns formed by multiple acoustic feature frames are overlooked. As a result, each feature frame needs to combine enough context from the original speech waveform in order to well represent a word-like unit. However, such an assumption fails to hold for acoustic features like MFCCs even bottleneck features fail to represent all of the context necessary to identify a word unless they are specifically trained to do so. Without sufficient context to represent a complete word, the mixture assumption assigns the phone sequences of any order to equal probability. In the phone-level discovery case, the word "god" and "dog" will be equivalent. One natural approach to capture multi-frame patterns is to replace the framewise model with an acoustic model that operates at the level of segment or a sequence of audio frames [3]. A *segment* is defined in this context as a variable-length sequence of consecutive audio feature frames that potentially represent a word or subword unit. Models of segment probabilities are called *segmental models*. The difference between segmental models and frame-wise models is illustrated on Fig. (3).

A. Phone-level word discovery

Models that represent the likelihood of each frame and the frame sequence with separate probabilities are called sequence models. On of the simplest types of sequence model represents the dependence between alignments. We may replace the uniform prior assumption with the following *Markov assumptions*: $p(i(t)|i(1:t),T_y) =: p(i(t)|i(t-1),T_y)$. Eq. (5) then simplifies to:

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{(T_y+1)^{T_x}}\sum_{\mathbf{A}\in\mathcal{A}}\prod_{t=1}^{T_x} p(i(t)|i(t-1),T_y)p(x_t|y_{i(t)}), \tag{20}$$

where $p(i(1)|i(0)) = p(i(1))$. This model, as first shown by [36], [37], is a hidden Markov model (HMM) with alignment vectors as the states. Maximizing Eq. (20) amounts to collecting the expected counts:

$$\frac{\langle c(i(t-1),i(t)|\mathbf{x},\mathbf{y})\rangle}{\sum_{i,j=1}^{T_y}\langle c(i,j|\mathbf{x},\mathbf{y})\rangle} = \frac{\alpha_{t-1}(i(t-1))p(i(t)|i(t-1),T_y)p(x_t|y_{i(t)})\beta_t(i(t))}{\sum_{i=1}^{T_y}\alpha_{t-1}(i(t-1))p(i|i(t-1),T_y)p(x_t|y_i)\beta_t(i)} \tag{21}$$

$$\langle c(x_t|y_i;\mathbf{x},\mathbf{y})\rangle = \sum_{i'=1}^{T_y}\langle c_t(i',i|\mathbf{x},\mathbf{y})\rangle, \tag{22}$$

where $\alpha_t(i) := p(x_{1:t}, i(t) = i|\mathbf{y})$ and $\beta_t(i) := p(x_{t+1:T_x}|i(t) = i,\mathbf{y})$ can be updated iteratively via dynamic programming.

Standard Viterbi decoding can be applied to find the optimal alignment between concept and phone:

$$i^*(t) = \max_{1\le i\le T_y}\{p(\mathbf{x}_{1:t-1}, i(1:t-1)|\mathbf{y})p(i|i(t-1))p(x_t|y_i)\}. \tag{23}$$

For image retrieval, the translation probability from Eq. (20) is used to find the optimal set of image concepts:

$$\mathbf{y}^* = \max_{\mathbf{y}} p(\mathbf{x}|\mathbf{y}) = \max_{\mathbf{y}}\sum_{i=1}^{T_y}\alpha_{T_x}(i).$$

B. Audio-level word discovery

1) Static Subword-level Segmentation Approach: While modeling intermediate correlations between alignments may be sufficient for phone-level word discovery, the audio-level feature may require modeling directly on the level of the segment. Assume that the first word starts at the first frame and the last word ends at the last frame with no gap or overlap between segments. Let N be the number of segments, and we have $s_1 \equiv 1$, $s_j < s_{j+1}, 1 \le j \le N$ and $s_{N+1} \equiv T_x + 1$.

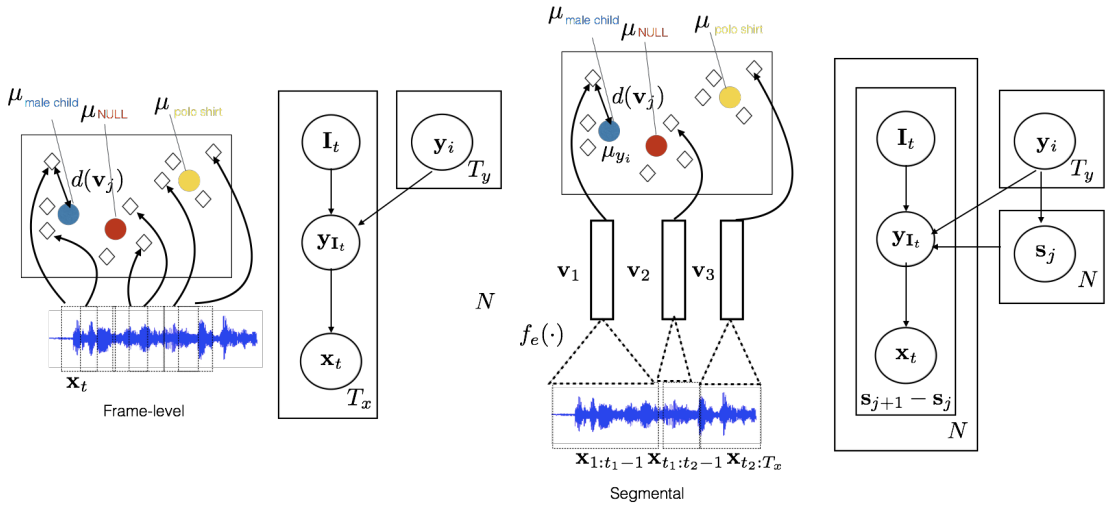

Fig. 3. Comparison between frame-wise models and segmental models. Frame-wise models cluster each acoustic frame directly, without taking into account contextual information on the segment level; in contrary, segmental models exploits segmental information by clustering fixed length embedding representations of the segments, the boundaries of which are learned through the hidden factors s_j, $j = 1, \cdots, N$.

A *segmentation* is then uniquely characterized by the vector $\mathbf{s} = [s_1, \cdots, s_N]$, where s_j is the starting frame of segment j. The segmentation can be alternatively represented by a binary *boundary vector* $\mathbf{b} \in \{0,1\}^{T_x}$ such that $b_{s_j} = 1, j = 1, \ldots, N$ and 0 otherwise. Let $\mathcal{S}_{T_x,N}$ denote all the segmentations of length N for feature sequence of length T_x. By the definition of the segment, all feature frames belonging to the same segment will align to the same image concept. As a result, if the segmentation is known, the alignment matrix $\mathbf{A}$ can be compressed into a smaller matrix $\tilde{\mathbf{A}} \in \{0,1\}^{T_y \times N}$ with the property in Eq. (1), which will be referred to as the *assignment matrix*. Let the set of all assignment matrices for N segment be $\tilde{\mathcal{A}}_N$: the translation probability of the segmental model can be expressed as:

$$p(\mathbf{x}|\mathbf{y}) = \sum_{N=N_{min}}^{N_{max}} \sum_{\mathbf{s}\in\mathcal{S}_N} p(\mathbf{s}|\mathbf{y}) \sum_{\tilde{\mathbf{A}}\in\tilde{\mathcal{A}}_N} p(\tilde{\mathbf{A}}|\mathbf{s},\mathbf{y}) p(\mathbf{x}|\tilde{\mathbf{A}},\mathbf{s},\mathbf{y}). \tag{24}$$

The expression shows that once the segmentation is fixed, the problem reduces to maximizing the likelihood of the segments given the image concept.

Suppose there is a subword-level segmentation $\mathbf{s}^*$ based on prior knowledge, for instance, the acoustic properties of the syllable units, the problem reduces to maximizing:

$$p(\mathbf{x}|\mathbf{y}) \approx p(\mathbf{x}|\mathbf{y},\mathbf{s}^*). \tag{25}$$

Consequently, with a similar Markov assumption as in the phone-level word discovery, the translation probability can be broken into probabilities of the subword units:

$$p(\mathbf{x}|\mathbf{y},\mathbf{s}^*) = \frac{\epsilon}{(T_y+1)^N} \prod_{j=1}^{N} \sum_{i(j)=0}^{T_y+1} p(i(j)|i(j-1),\mathbf{y}) p(x_{s_j:s_{j+1}}|y_{i(j)}, m(j)), \tag{26}$$

which can be modelled using an HMM with a mixture observation density: $p(x_{s_j:s_{j+1}}|y_{i(j)}) = p(m(j)|y_{i(j)},\mathbf{s}) p(x_{s_j:s_{j+1}}|y_{i(j)}, m(j))$. The Baum auxiliary function for the segmental model is then:

$$Q(\bar{\Theta},\Theta) = \sum_{j=1}^{N} \mathbb{E}_{\mathbf{i},\mathbf{m}}[\log p(x_{s_j^*:s_{j+1}^*}, i(j), m(j)|i(j-1), y_{i(j)}, \Theta)|\mathbf{x},\mathbf{y},\bar{\Theta}]. \tag{27}$$

The challenge here is that the segments are variable in length and $p(x_{s_j:s_{j+1}}|y_{i(j)})$ can not be directly modelled with a fixed dimensional density function. Besides modeling each segment using an HMM [5], one existing approach [3] is to embed each segment into a fixed-dimensional space as an *embedding vector* so the embedding vector can be modelled with a fixed dimensional density function. Denote the embedded j-th segment as $\tilde{x}_j = f_e(x_{s_{j-1}:s_j})$, $s_0 := 1$ and $\tilde{\mathbf{x}} := [\tilde{x}_1, \cdots, \tilde{x}_N]$, the assumption amounts to $p(x_{s_j:s_{j+1}}|y_{i(j)}) p(\tilde{x}_j|y_{i(j)})^{s_{j+1}-s_j}, p(\tilde{x}_j|y_{i(j)}) = \sum_{m=1}^{M} c_m(y_{i(j)}) \mathcal{N}(\tilde{x}_j|\mu_m(y_{i(j)}), \Sigma_m(y_{i(j)}))$. A special case of such embedding assumptions is that most spoken words are distinguisable by the ear even after they are properly scaled to the same size. Therefore, the embedding vector can simply be the resampled version of the segment it represents. While duration carries information in many languages and the embedded segment may suffer from loss of high-frequency information when the downsampling rate is too high, embedding segments to fixed length greatly reduces the complexity of the model. This model will be referred to as the *static segmental HMM*.

2) Static Word-level Segmentation Approach: Suppose $\mathbf{s}^*$ is instead a word-level pre-segmentation and the pseudo-word segments satisfy the following *segmental mixture assumption*: Given the assignment, each segment depends only on its aligned image concept, thus $p(\mathbf{x}_{s_j:s_{j+1}-1}|\mathbf{x}^-, \tilde{\mathbf{A}}, \mathbf{s}, \mathbf{y}) = p(\mathbf{x}_{s_j:s_{j+1}-1}|\mathbf{y}_{i(s_j)}) = p(\tilde{\mathbf{x}}_j|\mathbf{y}_{i(s_j)})^{s_{j+1}-s_j}$, where $\mathbf{x}^-$ denotes the segments other than j. Combined with the embedding vector approach, a simplified or enriched mixture model can be used to model the segments. The key difference between

the word-level segmentation approach to the subword-level approach is the omission of the transition probabilities for model simplicity, as the intermediate contextual information between words is much weaker than those between image concepts and word units.

3) Dynamic Segmentation Approach: The approaches above fix the noisy word boundaries through the clustering process, and their performance is therefore constrained by the quality of the pre-segmentation. It is more appealing to jointly refine the segmentation and clusters during training [3]. Suppose, given the segmentation, the utterance can be modeled by an enriched mixture model, then the Baum auxiliary function is:

$$Q(\bar{\Theta},\Theta) = \sum_{N=N_{min}}^{N_{max}} \sum_{\mathbf{s}\in\mathcal{S}^N} p(\mathbf{s}|\mathbf{x},\mathbf{y},\bar{\Theta}) \mathbb{E}_{\tilde{\mathbf{A}},\mathbf{m}}\left[\log p(\mathbf{x},\mathbf{s},\mathbf{m},\tilde{\mathbf{A}}|\mathbf{y},\Theta)|\mathbf{x},\mathbf{y},\mathbf{s},\bar{\Theta}\right]. \quad (28)$$

One key challenge of the jointly segment-and-cluster approach is to model the distribution of the segmentation $p(\mathbf{s}|\mathbf{x},\mathbf{y})$ and $p(\mathbf{s}|\mathbf{y})$. Similar to the prior probabilities of alignments, the summation over the priors and posteriors of segmentation in Eq. (24) and Eq. (28) respectively has $(N_{max} - N_{min})\begin{pmatrix} T_x \\ N-1 \end{pmatrix}$ terms and requires a prohibitive amount of parameters. Suppose $p(\mathbf{s}|\mathbf{x},\mathbf{y})$ and $p(\mathbf{s}|\mathbf{y})$ are known, we can approximate the sums by its unbiased sampled version:

$$p(\mathbf{x}|\mathbf{y}) \approx p(\mathbf{x}|\mathbf{y},\mathbf{s}') \quad (29)$$

$$Q(\bar{\Theta},\Theta) \approx \mathbb{E}_{\tilde{\mathbf{A}},\mathbf{m}}\left[\log p(\tilde{\mathbf{x}},\mathbf{s}'',\tilde{\mathbf{A}},\mathbf{m}|\mathbf{y},\Theta)|\mathbf{x},\mathbf{y},\mathbf{s}'',\bar{\Theta}\right] \quad (30)$$

$$=: Q(\hat{\Theta},\Theta|\mathbf{s}''), \quad (31)$$

where $\mathbf{s}' \sim p(\mathbf{s}|\mathbf{y}), \mathbf{s}'' \sim p(\mathbf{s}|\mathbf{x},\mathbf{y})$. This representation is unbiased since:

$$\mathbb{E}_{s'}[p(\mathbf{x}|\mathbf{y},\mathbf{s}')] = \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{y})p(\mathbf{x}|\mathbf{y},\mathbf{s}') = p(\mathbf{x}|\mathbf{y})$$

$$\mathbb{E}_{s''}[Q(\hat{\Theta},\Theta|\mathbf{s}'')|\mathbf{x},\mathbf{y},\bar{\Theta}]$$
$$= \sum_{\mathbf{s}} p(\mathbf{s}|\mathbf{x},\mathbf{y}) \sum_{\tilde{\mathbf{A}},\mathbf{m}} p(\mathbf{m},\tilde{\mathbf{A}}|\mathbf{s},\mathbf{x},\mathbf{y}) \log p(\tilde{\mathbf{x}},\mathbf{s}'',\mathbf{m},\tilde{\mathbf{A}}|\mathbf{y},\Theta)$$
$$= Q(\bar{\Theta},\Theta)$$

Notice that this expressions are the same as those of the models with fixed segmentation, except the segmentations are now drawn randomly rather than deterministically.

To model $p(\mathbf{s}|\mathbf{x},\mathbf{y})$ and $p(\mathbf{s}|\mathbf{y})$, One approach is to use *adaptor grammar* [5], [8], which assumes the segmentation boundary vector $\mathbf{b}$ to be a sequence of i.i.d Bernoulli random variables with parameter α_b. α_b is assumed to be generated by a symmetric Dirichlet process to encourage sparsity of the boundary vector. During inference, the boundary vector and Bernoulli parameter can be sampled efficiently using a collapsed Gibbs sampler. Once the boundary vector is fixed, each segment can be modelled separately. However, the approach is not scalable to large dataset on the audio level since it requires intensive amount of Gibbs sampling at every time step. The second approach is employed by the Bayesian embedded segmental GMM (BESGMM) system [3] and assumes each segmentation s_j to be uniformly distributed between $[0, s_{j+1}]$. Under this assumption and the segmental mixture assumption, $\mathbf{s}$ can be sampled backward from the posterior $p(\mathbf{s}|\mathbf{x},\mathbf{y})$:

$$\begin{aligned} s_j &\sim p(s_j|\mathbf{s}_{-j},\mathbf{x},\mathbf{y}) \\ &\propto p(s_j, s_{j+1}, \mathbf{x}_{1:s_{j+1}}|\mathbf{y}) \\ &\propto p(s_{j+1}|\mathbf{y})p(\mathbf{x}_{1:s_{j+1}}|s_j, s_{j+1},\mathbf{y}) \\ &\propto p(\mathbf{x}_{1:s_j}|\mathbf{y})p(\tilde{x}_j|\mathbf{y})^{(s_{j+1}-s_j)}, \end{aligned}$$

where the second $\propto$ uses the uniform assumption of s_j. $p(\mathbf{x}_{1:s_j}|\mathbf{y})$ can be computed recursively with dynamic programming, and the last $\propto$ comes from the segmental mixture assumption. The key motivation of the segmental mixture assumption is to prevent both under- and over-segmentation: if instead $p(\mathbf{x}_{1:s_j}|\mathbf{y}) = \prod_{t=s_j}^{s_{j+1}} p(\mathbf{x}_t|\mathbf{y})$, the probabilities will tend to decay as the sequence becomes longer; if $p(\mathbf{x}_{1:s_j}|\mathbf{y}) := p(\tilde{x}_j|\mathbf{y})$, the model will achieve the highest likelihood if it treats the whole sentence as the segment. Further, the expression also models the dependence between the adjacent segments with the term $p(\mathbf{x}_{1:s_j}|\mathbf{y})$.

Optimizing Eq. (29) with the EM algorithm results in a formula in terms of the following expected count:

$$\frac{\langle c_j(i,m;\mathbf{s},\mathbf{x},\mathbf{y})\rangle}{\sum_{i'=1}^{T_y}\sum_{m'=1}^{M}\langle c_j(i',m';\mathbf{s},\mathbf{x},\mathbf{y})\rangle} \quad (32)$$

$$= \frac{(c_m(y_i)\mathcal{N}(\tilde{x}_j|\mu_m(y_i),\Sigma_m(y_i)))^{s_{j+1}-s_j}}{\sum_{i=0}^{T_y}\sum_{m=1}^{M}(c_m(y_i)\mathcal{N}(\tilde{x}_j|\mu_m(y_i),\Sigma_m(y_i)))^{s_{j+1}-s_j}}. \quad (33)$$

The model infers the optimal segmentation and alignment using Viterbi decoding:

$$s_j^*, i^*(j) = \max_{s_j,i}\{p(\mathbf{x}_{1:s_j}|y_i)p(\tilde{x}_j|y_i)^{s_{j+1}-s_j}\}, \quad (34)$$

where similar to the sampling step, the $p(\mathbf{x}_{1:s_j}|y_i)$ is evaluated forward from $s_j = 1$ while the optimization starts from backward from $s_{j+1} = T_x + 1$. This model is based on BES-GMM [3] with additional image concepts input and referred later as dynamic segmental model. When the segmentation is fixed during training, the model is referred as static segmental GMM. [3] also proposes a simplified version of the BES - GMM called ES-KMeans, which can be naturally adapted to the multimodal setting to optimize:

$$\min_{\mathbf{s},\mathbf{A},\mathbf{m}} \sum_{j=1}^{N}(s_j - s_{j-1})\|v_j - \mu_{m(j)}(y_i(j))\|_2^2. \quad (35)$$

This simplified model is referred to later as the simplified dynamic segmental model.

VII. NMA MODELS

Similar to the SMA model, the NMA makes use of several assumptions:

1) *Dominant path assumption*: There is a "dominant" alignment $\mathbf{A}^*$ such that $p(\mathbf{A}^*|\mathbf{x}) \approx 1$;
2) *Embedding assumption*: The input representation of the phone x_t can be compressed into a lower-dimensional

embedding $h_t(x_t) =: \mathbf{h}_t$, $\forall t \in \{1, \cdots, T_x\}$. Let $\mathbf{h} = [\mathbf{h}_1, \cdots, \mathbf{h}_{T_x}]$;

3) There exists some "context" vector $\mathbf{c}_i(\mathbf{h}, A^*), i = \{1, \cdots, T_y\}$ such that y_i is conditionally independent of $\mathbf{h}$ given $\mathbf{c}_i$.
4) There exists some "state" vector $\mathbf{s}_i(\mathbf{y}_{1:i-1})$ such that y_i is independent of $\mathbf{y}_{1:i-2}$ given $\mathbf{s}_i$ and y_{i-1}.

The dominant path assumption is in stark contrast to the uniform prior assumption of the SMA model and instead of summing over all possible alignments, only a single soft alignment is used to compute the probability.

By sequentially applying the assumptions above, Eq. (7) is then simplified to one term:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{A}) = \prod_{i=1}^{T_y} p(y_i|y_{1:i-1}, \mathbf{x}, \mathbf{A}^*) = \prod_{i=1}^{T_y} p(y_i|\mathbf{y}_{1:i-1}, \mathbf{h}, \mathbf{A}^*) \tag{36}$$

$$= \prod_{i=1}^{T_y} p(y_i|\mathbf{y}_{1:i-1}, \mathbf{c}_i) \tag{37}$$

$$= \prod_{i=1}^{T_y} p(y_i|y_{i-1}, \mathbf{s}_i, \mathbf{c}_i). \tag{38}$$

In the standard attention model [15], the dominant alignment is learned via a soft alignment:

$$A_{it} := \alpha_{it} = \frac{\exp(e(\mathbf{h}(\mathbf{x}_t), \mathbf{y}_i)/T)}{\sum_{j=1}^{T_y} \exp(e(\mathbf{h}(\mathbf{x}_t), \mathbf{y}_j)/T)} \tag{39}$$

where $\mathbf{e}(\cdot)$ can be learned by a feedforward neural network. Compared to the original attention mechanism, our attention has two main differences: First, the attention weights are normalized across concepts for a given acoustic feature frame; second, the decoder state is not fed into the attention to avoid dependency on future states when computing the softmax over all the image concepts. These assumptions are necessary to represent $p(\mathbf{y}|\mathbf{x}, \mathbf{A})$ in a problem like ours, with no fixed sequencing of the concepts in $\mathbf{y}$, and these assumptions may be ignored if the network learns $p(\mathbf{x}|\mathbf{y}, \mathbf{A})$ instead. However, such modeling choice is at odds with the discriminative nature of the NMA model: the acoustic features can be continuous and cannot be divided into a finite number of categories, and thus the output probability cannot be modelled by a softmax function. The context vectors are then learned by:

$$\mathbf{c}_i = \sum_{t=1}^{T_x} \alpha_{it}^* \mathbf{h}_t, \tag{40}$$

rather than being normalized across feature frames for a given concept; the assumption (3) of this section can be viewed as a soft version of the SMA assumption (3) in the previous section: in the SMA model, concept y_i depends only on the phones that align to it, which is equivalent to the special case when α_{it}^* in Eq. (40) is either 0 or 1. Therefore, the set of probabilities $\{p(y_i|y_{i-1}, \mathbf{s}_{i-1}, \mathbf{c}_i)\}_{i=1}^{T_y}$ can be learned using a recurrent neural net f with state vector s_i. The problem then boils down to learning the functions h, e, f such that the log-likelihood of the concepts given the phone sequence is maximized.

VIII. Experiments

A. Datasets

Our dataset consists of 7996 images that are present in the Flickr8k [17] and Flickr30k corpus [38]. We choose to use Flickr8k primarily because it contains phrases that describe particular regions of the image, for instance, "a girl in a pink dress". The phrase-level segmentation is used as our gold alignment. We only used images that appeared in the Flickr8k dataset so that we can compare our results to the speech-to-image [21] and text-to-image [20] system. In order to make sure that we used the same dataset as in [20], [21], we divided the data into training set (6996 images) and test set (the same 1000 images used in [20], [21]). We use Flickr30kEntities [39] to extract image concepts from the phrase-level descriptions and merged similar entities using WordNet [40] synsets. By considering captions in which every concepts appear at least 10 times in the training and test sets, we are able to find a list of 1547 concepts. Repeated concepts in a sentence are merged to maintain the many-to-one mapping between the phone sequence and the image concepts. To generate the phone sequence for the caption, we used the CMU dictionary which consists of 39 distinct phone labels and 69 tokens in total with the stress symbols. Words without an entry in the dictionary are simply replaced with an `UNK` symbol. For audio-level discovery, we used audio from the Flickr-audio dataset [21], which are spoken captions for Flikcr8k collected on Amazon Mechanical Turk. We extract the MFCC features as in [2], [26] with 12 cepstral coefficients, 25 ms window and 10 ms overlaps, and normalize them across speakers with cepstral mean and variance normalization (CMVN). Another acoustic feature we use is the multilingual bottleneck feature from the BUT project trained on 20 different languages [32]. Utterances longer than 2000 frames are truncated. Using word-level forced alignment by an ASR system [41], we are able to convert the phone-level gold alignment to the alignment of acoustic feature frames. Many of the audio waveforms are corrupted and did not have successful forced alignments, so we filter them out from our dataset, and leaving with 6610 images.

B. Evaluation metrics

For the word-discovery task, we evaluate our systems by comparing the predicted and gold alignments. The *alignment accuracy* is defined as the percentage of phones that align to the correct image concept, which is equal to the complement of the alignment error rate (AER) metric commonly used in machine translation systems (see, for example, [36]). In addition, we use the alignment *recall*, *precision* and *F-score* to evaluate the quality of the alignment. The concept-specific *alignment recall* is defined as the percentage of correctly aligned phones or audio frames to a given concept over all the phones or audio frames that should be aligned to the concept; the concept-specific *alignment precision* is the percentage of correctly aligned phones or audio frames to a given concept over all the predicted aligned phones or audio frames to the concept. The corpus-level alignment precision and recall are the average of their concept-specific counterpart across all

image-caption pairs. The F-score is then the harmonic average between the *alignment recall* and the *alignment precision*. For visualization purposes, we also computed the concept-specific alignment F-score over the entire corpus for each concept. Besides, we also used the grouping, boundary and token/type metrics used in 2017 ZRSC [7]. Since our dataset is not balanced (for example, the `NULL` concepts are much more common than any other concepts) and the model can achieve high alignment accuracy by aligning only to the most frequent concepts, such retrieval-based metrics help us to fairly evaluate our systems. For the speech-to-image retrieval task, we follow [20], [21] to use recall@1, 5, 10 to measure the performance of our system. We assume one-to-one mappings between image and caption, despite having a large number of image-caption pairs with similar concepts. To compare the performance, we use the speech-to-image system by [21] and the text-to-image system by [20] as baselines.

C. Model parameters

The main parameters of the phone-level SMT are the translation probabilities and alignment probabilities. We run the simplified mixture model until convergence and the enriched and segmental models for 50 iterations. We initialize $p(x_t|y_i)$ by adding one to the numerator when phone x_t and y_i co-occur in a sentence. For the segmental model, we initialize the alignment transition matrix with a uniform distribution for each row.

For the NMT system, we used a 512-dimensional embedding layer followed by a single layer bidirectional LSTM with 512 hidden nodes as the encoder and another single layer LSTM with 512 hidden nodes as decoder.

For the audio-level simplified mixture model, the main hyperparameter is the number of mixtures per concept M and the dimension of the acoustic embedding. We found that $M = 5$ produces the best results and any value larger than 5 leads to many empty clusters. For the enriched mixture model, the covariance matrices become singular as the posterior of any certain mixture becomes too small, so we simply fix the covariance matrices to be diagonal with diagonal entries equal to 0.002 for the framewise approach and 2 for the pre-segmental approach, based on the increase of log-likelihood during training. Both models are initialized with randomly generated clusters and assignments. For the segmental model, we used $M = 1$ for computational efficiency and initialized the cluster randomly. For the embedding dimension, we experimented with 10 and 20 equally spaced frames and found only marginal improvement when using 20 equally spaced frames with proper resampling or interpolation, so we follow [3] to concatenate 10 equally spaced frames. Again following [26], we pre-segmented the audio to syllable-level using the unsupervised syllable-segmentation system by [2] and only consider the boundaries detected by the system during the segment step. This limits the performance of our system to the coverage of their system, but significantly reduces the disk space for storing the embeddings and inference time. To ensure a better coverage of the true word boundaries, we combine the landmarks detected by all three syllable segmentation algorithms (Oscillator, EnvMin, VSeg) [2] and remove boundaries that are within 20 ms from each other. This results in a coverage of about 83.5% of all the word boundaries to be within 30 ms. Due to run time constraint, we experiment mainly with the ES-Kmeans based segmental system and leaves the comparison with BES-GMM for future work. From our preliminary experiments, the word segmentation results of the two models are very similar.

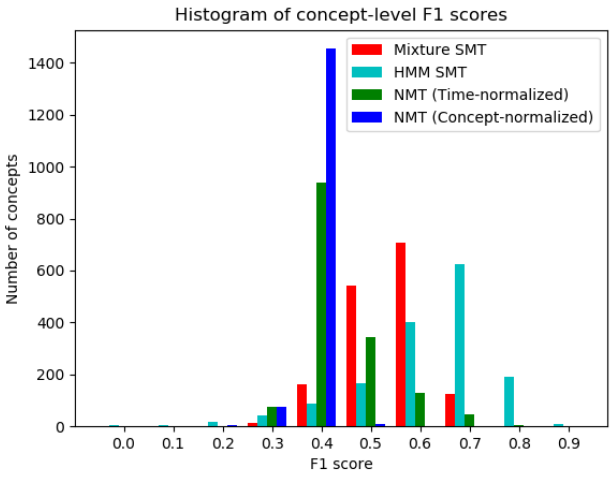

Fig. 4. F1 histograms for phone-level models. The horizontal axis is the cutoff values of concept-specific F-scores for each bar and the vertical axis represents the number of concepts with alignment F-scores that lie in between the cutoff F-scores. There are 1547 concepts in total and the more the F-score distribution leans toward F-score=1, the better the model performs.

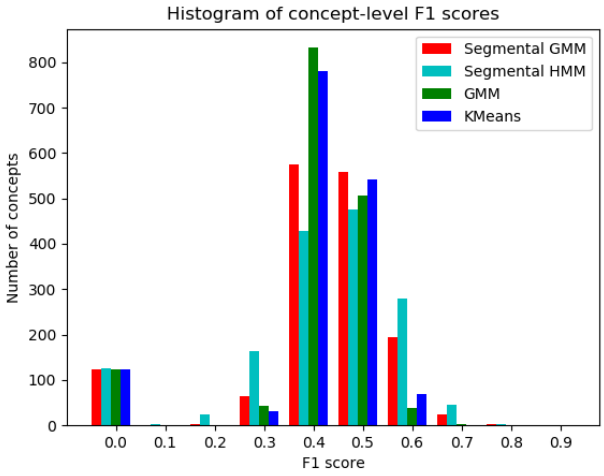

Fig. 5. F1 histograms for audio-level models with MFCC features. F1 histograms for phone-level models. The horizontal axis is the cutoff values of concept-specific F-scores for each bar and the vertical axis represents the number of concepts with alignment F-scores that lie in between the cutoff F-scores. There are 1547 concepts in total and the more the F-score distribution leans toward F-score=1, the better the model performs.

D. Results

1) Phone-level discovery: The phone-level word discovery results are summarized in Table (I). From the retrieval based

		Simplified Mixture	Enriched Mixture	Segmental (HMM)	NMT	Adaptor Grammar	Majority
Alignment	Accuracy	43.8	40.0	**55.4**	41.5	-	47.8
	Precision	46.7	41.4	**56.7**	33.0	-	26.8
	Recall	52.9	46.9	**67.4**	29.2	-	13.4
	F1	49.6	44.0	**61.6**	31.2	-	17.8
Grouping	Precision	**65.5**	54.3	33.2	38.7	-	0.01
	Recall	**76.9**	67.2	70.0	54.5	-	1.0
	F1	**70.7**	60.1	45.0	45.3	-	0.02
Boundary	Precision	12.6	9.6	**35.1**	9.38	32.01	1.0
	Recall	62.8	47.2	46.0	15.7	**74.0**	30.0
	F1	21.0	16.0	39.8	11.7	**44.7**	46.15
Token	Precision	0.508	0.199	5.72	0.153	**12.05**	0.05
	Recall	2.24	1.35	11.7	1.09	**25.16**	0.009
	F1	0.993	0.275	7.68	0.268	**16.3**	0.015
Type	Precision	3.72	1.51	3.37	0.296	**29.49**	0.05
	Recall	13.5	11.1	**51.7**	3.82	4.31	0.009
	F1	5.83	2.66	6.33	0.549	**7.53**	0.015
Avg. Word Length (True Avg. = 9.13)		2.58	2.62	**6.88**	20.7	5.77	-

TABLE I
PHONE-LEVEL WORD DISCOVERY RESULTS INCLUDING THE ALIGNMENT ACCUARCY, ALIGNMENT/GROUPING/BOUNDARY/TOKEN/TYPE RECALL, PRECISION, F-SCORE (ALL IN %) AND THE AVERAGE WORD LENGTH (IN NUMBER OF PHONES) ARE SHOWN

	10 Highest	10 Lowest
Simplified Mixture	ocean, sunglasses, paper, adolescent, flower, jean, water, people, baseball_glove, water_scooter	instrument, expression, hole, seat, NULL, lawn, drive, rug, fabric, slope
Segmental (HMM)	sunglasses, motorcyclist, resort_area, adolescent, microphone, couple, people, cyclist, ocean, girl	NULL, base, seat, headscarf, bride, reversal, lawn, fabric, log, line
NMT (Normalized Over Time)	man, male_child, goggles, plant, plaything, grass, hair, guitar, sweatshirt, shirt	fountain, truck, puddle, puppy, adolescent, court, cat, snowboarder, slide, climber

TABLE II
10 CONCEPTS WITH THE HIGHEST AND LOWEST ALIGNMENT F1 SCORES FOR DIFFERENT PHONE-LEVEL WORD DISCOVERY MODELS

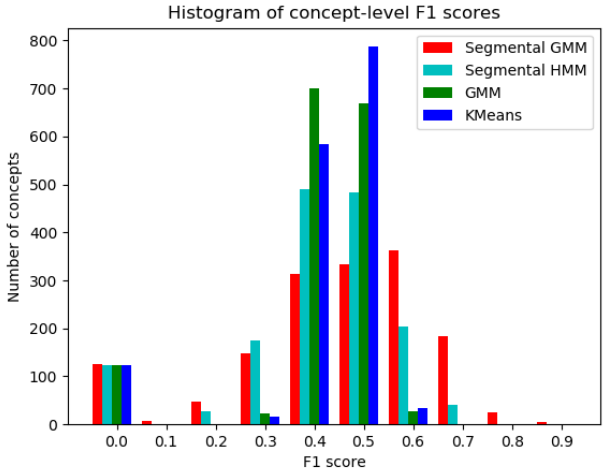

Fig. 6. F1 histograms for audio-level models with bottleneck features. F1 histograms for phone-level models. The horizontal axis is the cutoff values of concept-specific F-scores for each bar and the vertical axis represents the number of concepts with alignment F-scores that lie in between the cutoff F-scores. There are 1547 concepts in total and the more the F-score distribution leans toward F-score=1, the better the model performs.

scores, we see that segmental model is better at aligning phone to image concepts as it performs better in both recall and precision than the simplified model. The enriched model, however, performs worse than the simplified model potentially because the image concepts do not appear in approximately the same order as the corresponding words in the utterances. Further, we compare our models with the adaptor grammar [8], which only has access to the ground truth phone labels. We see that the SMT models perform worse than the adaptor grammar in the boundary, token and type F1 score, but the segmental models perform slightly better in the boundary precision score and much better in the type recall score. This can be explained by the fact that the phone-level segmental model does not use word-level contextual information as the adaptor grammar does, and tends to discover many incorrect sub-strings for each word type, which lowers its type and token precision scores. Since the adaptor grammar can be viewed as a more thorough-going segmental model than the HMM segmental model, this result confirms again the superiority of the segmental model over the frame-wise model.

The alignment probability/attention matrices for the four models are shown in Fig. (7). For the segmental model, the alignment probabilities are defined as the Viterbi probabilities normalized across concepts at a given time. From the alignment probability, we notice that the segmental model has much sparser alignment probabilities and produces much more continuous pseudo-words that are closer the length of a real word, suggesting that contextual information between phones is crucial for discovering word-like units. However, the segmental model seems to be worse at distinguishing `NULL` from actual concepts, making many false positives, indicating that the `NULL` concept is statistically distinct from the rest of the concepts and should be modelled separately. Another cause of the false discovery stems from the limitation of a concept-independent jump transition probabilities: it appears

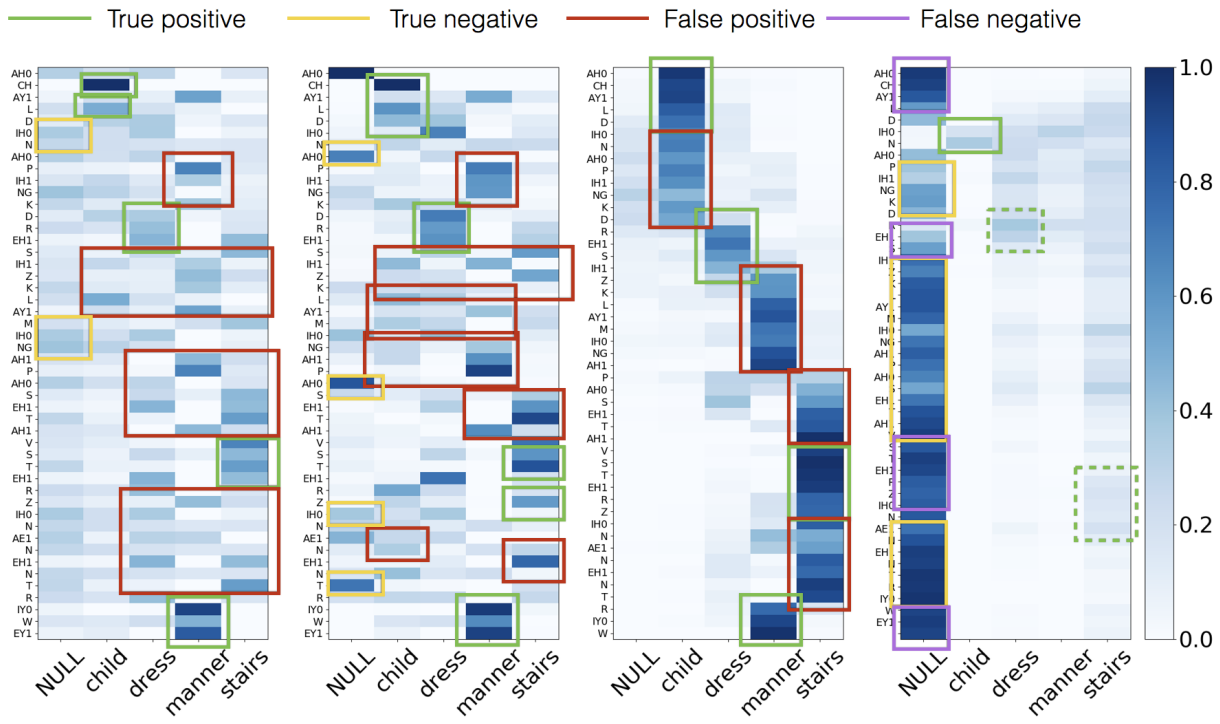

Fig. 7. Attention/alignment probability matrices annotated with discovered word-like units of various phone-level models for the utterance "A child in a pink dress is climbing pass the stairs of a entry way". Leftmost: simplified mixture model; middle left: enriched mixture model; middle right: segmental model; rightmost: NMT

that the segmental model prefers smaller jumps than larger ones, causing a exceedingly slow transition from one concept to another. This is evident from the example attention plot: almost every transition moves one step at a time. In particular, we can clearly see that the model stays at the concept "child" for eight extra phones until the alignment probability drops gradually. Another observation is that the accuracy of SMT is generally lower than the retrieval metrics as opposed to higher in the case of NMT primary because SMT is less biased towards concepts that appear more frequently while NMT tends to memorize the prior distribution of the concepts in the training data.

The histograms of concept-specific F1 scores for the mixture, segmental and NMT models are shown in Fig. (4). The mixture SMT and HMM SMT both outperforms the the NMT models with a large gap between their F1 score distributions. While the mean F1 scores of the mixture and HMM SMT are centered around 0.6 and 0.7 respectively, the mean NMT score stays at 0.5. Further, the F1 score of the segmental model for 1200 out of 1547 concepts is higher than 60%, followed by 900 concepts for the mixture model and 200 for the NMT model. This result suggests that the overall performance of the segmental model is superior to the other models.

The top 10 easiest and hardest concepts for different models are shown in Table II. As we can see, both the phone-level SMT and NMT models are better at discovering human concepts such as "adolescent" and "people" in the case of SMT and "man" and "male_child" in the case of NMT. The models are also able to discover some human-related concepts such as "sunglasses" and "goggles". Shorter, more concrete words with one or two syllables such as "grass" and "ocean" tend to be easier to discover than longer, more abstract words such as "expression" and "instrument".

		SylSeg	ES-KMeans	Simplified Dynamic Segmental
Recall	MFCC	83.5	35.7	35.6
	BN		-	38.2
Precision	MFCC	33.9	41.0	40.2
	BN		-	41.6
F1	MFCC	48.2	38.1	37.2
	BN		-	39.9

TABLE III
AUDIO-LEVEL WORD SEGMENTATION BOUNDARY RECALL, PRECISION AND F-SCORE (ALL IN %)

Unsurprisingly, the segmental model performs better at discovering longer words such as "microphone" and "motorcyclist" than the mixture model, and the sets of easy and hard concepts for SMT and NMT models can be quite different. One example is the concept "adolescent" is among the easiest concepts for SMT models but one of the hardest concepts for NMT. Another example is the placeholder `NULL` concept, which is among the hardest concepts for SMT to discover but does not appear in the top 10 hardest concepts of NMT.

2) Audio-level discovery: Table (IV) shows the audio-level word discovery results. Segmental models generally outperform the simplified model approaches and produce words more realistic lengths. The static segmental models generally performs better than the dynamic model when using MFCC. The static segmental GMM performs best when using either

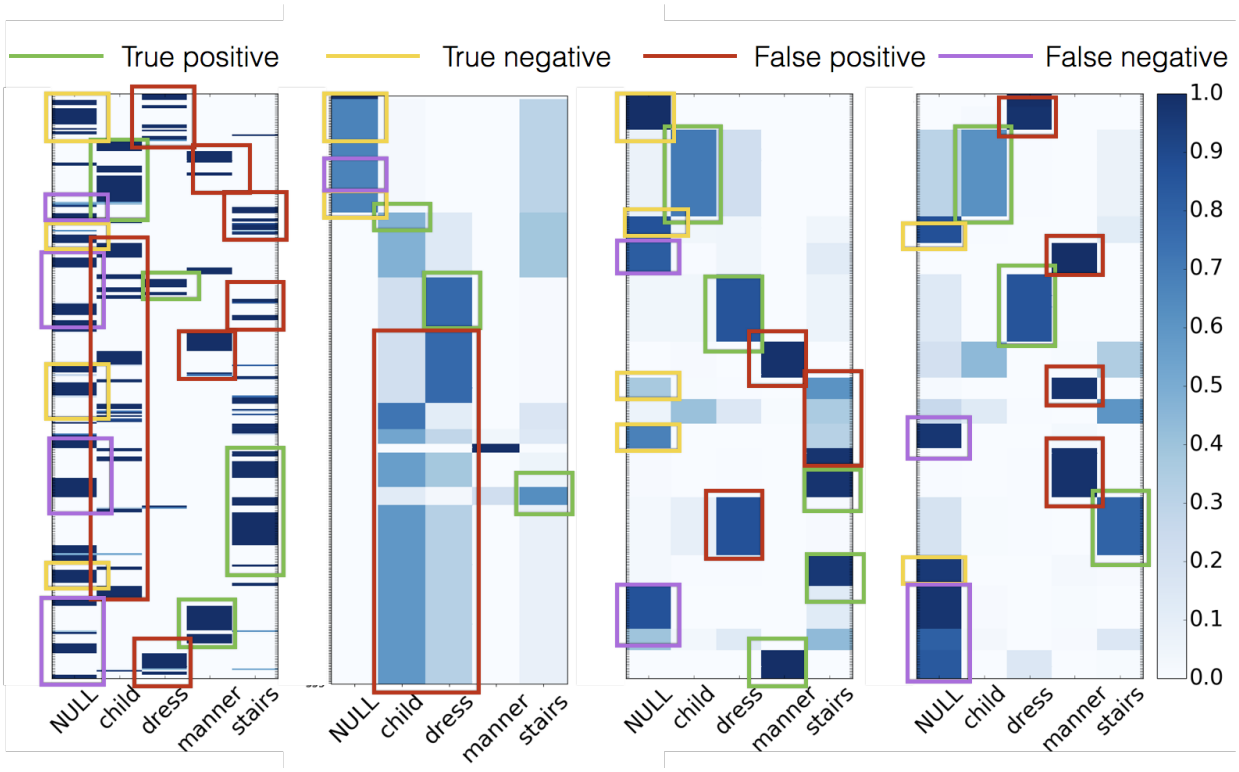

Fig. 8. Alignment probability matrices annotated with discovered word-like units of various audio-level models for the utterance "A child in a pink dress is climbing pass the stairs of a entry way". Leftmost: enriched frame-wise mixture model with BN feature; middle left: dynamic segmental KMeans; middle right: static segmental GMM; rightmost: static segmental HMM. The results of the frame-wise KMeans model is similar to the frame-wise mixture model and is omitted.

		Simplified Mixture	Enriched Mixture	Static Segmental GMM	Static Segmental HMM	Simplified Dynamic Segmental	NMA (Normalized ov. Concept)	Majority
Accuracy	MFCC	32.2	27.1	36.6	31.2	28.0	19.3	55.1
	BN	35.6	36.6	**40.8**	37.9	29.7	20.3	
Recall	MFCC	30.4	27.5	36.1	32.7	26.2	25.7	27.0
	BN	36.5	36.4	**44.1**	39.7	37.4	27.0	
Precision	MFCC	29.8	26.9	34.7	30.6	23.9	12.2	15.4
	BN	34.6	34.3	**40.5**	36.5	39.4	13.3	
F1	MFCC	30.1	27.2	35.4	31.6	25.0	23.1	19.6
	BN	35.5	35.3	**42.3**	38.0	38.3	17.8	
Avg. Word Length (True Avg. = 61.2)	MFCC	3.14	3.26	39.9	40.8	140.1	36.9	-
	BN	6.32	6.31	42.2	39.9	**71.8**	16.0	

TABLE IV
AUDIO-LEVEL WORD DISCOVERY ALIGNMENT ACCUARCY, RECALL, PRECISION, F-SCORE (ALL IN %) AND AVERAGE WORD LENGTH (IN NUMBER OF ACOUSTIC FRAMES)

MFCC or BN, suggesting that the dynamic segmentation may be unnecessary for our dataset since most words have one or two syllables. This is also consistent with the observation that the Markov assumption does not boost the performance of word discovery, suggesting that different syllables in the speech are weakly correlated. The dynamic segmental model produces words of the most realistic lengths when using BN but tends to under-segment and create extremely long segment when using MFCC. The bottleneck feature performs significantly better than the MFCC in all models possibly by incorporating more contextual information. Most models have lower F1 score than the most closely related phone-level models, e.g. SylSeg is worse than segmental and K-Means is worse than the simplified mixture models, possibly because of the aliasing introduced by the resampled embedding approach, the speaker variability and other losses of information during the feature extraction process. Indeed, from Table. (III), feature embedding seems to subvert the segmentation process since the feature-free SylSeg model has a higher F1 score than the feature-dependent ES-KMeans model and multimodal ES-Kmeans model, while the similar phone-level mixture models are likely to improve the F-score given a pre-segmentation.

The F1 histograms for audio-level word-discovery results are shown in Fig. (5)(6). The frame-wise approach has almost no concepts with F1-scores higher than 60%. The segmental approach performs better by having 600 concepts for the mixture model and 300 concepts for the Markov model with F-scores over 60%, though accompanying with about 100 con-

cepts with an F-score of 0. Within the segmental model, GMM has more concepts with higher F-scores than HMM possibly possibly because the correlation between segments is too weak to model meaningfully by the HMM. The large within-token variabilities of the acoustic units may also exacerbate the issue by making it hard to cluster units of the same type.

The alignment probability matrices for four audio-level models are shown in Fig. (8). Different from the phone-level plots, the plots are not generated by exponentiating the raw translation probabilities but as a smoothed version with some temperature T:

$$p'(i(t)|\mathbf{x},\mathbf{y}) = \frac{\exp(\log p(i(t)|\mathbf{x},\mathbf{y})/T)}{\sum_i \exp(\log p(i'|\mathbf{x},\mathbf{y})/T)}.$$

The smoothing does not alter the trend of the probabilities but makes it more visually informative. We found $T = 1000$ to produce figures with visually salient segmentations. Judging from the plots, the two static segmental models perform similarly in true discoveries and the HMM tend to discover more continuous words, but the HMM has more false positives and false negatives. Indeed, all except the static segmental HMM model cluster the first few frames of silence to `NULL` symbol. Overall, most models display a high level of false discoveries, possibly because the statistical properties of `NULL` symbol are different from other concepts in that it is more dependent on the other concepts present in an image. Further, the dramatic drop in F1 score indicates that the audio-level models are having troubles extracting the phonetic information from audio that is necessary to reduce the problem to the phone-level discovery.

3) Phone-to-concept retrieval: The result for phone-to-concept retrieval compared with the speech-to-image and text-to-image systems are shown in Table. (V). The performance of the simplified SMA retriever is just in between the speech-to-image system [21] and the text-to-image system by [20]. The segmental SMA retriever, however, performs better than both the speech-to-image and text-to-image models by about 300 %, 100 % and 75 % for recall@1, 5, 10 respectively relative to the text-to-image system. The gain in performance can be partly explained by the use of ground truth hard labels for the image concepts and phone labels for the caption used by the system and demostrates that proper representation of speech is a main challenge in speech-to-image system. It also demonstrates that detecting the entities is the key for the image retrieval task on our dataset. Nevertheless, many of the errors made by our systems come from either unknown image concepts or image with very similar top concepts, so there is a substantial room for improvement on the concept level, especially in making use of the contextual information between concepts in retrieval. For example, the concepts "shirt" and "hand" are more likely when a person is present. Indeed, the SMA-based retriever does not perform well in our preliminary experiments for the more challenging task of captioning primarily due to the lack of modeling of the relations between concepts.

IX. Conclusion

This paper describes a unified framework for multimodal word discovery with spoken captions and image concepts.

	Recall@1	Recall@5	Recall@10
Simplified SMT	9.42%	21.1%	29.1%
Segmental SMT	46.7%	65.2%	72.2%
Harwath&Glass [21]	-	-	17.9%
Karpathy [20]	10.3%	31.4%	42.5%

TABLE V
COMPARISON OF QUERY-BY-EXAMPLE IMAGE SEARCH WITH SPOKEN/PHONE SEQUENCE RESULT

Four systems were tested with three different representations of the spoken caption: the ground truth phonetic labels, MFCC and MBN. With the amount of data we have, the segmental SMT approach performs best in the phone-level while the enriched SMT performs best in the audio-level, according to our evaluation metrics. We applied our word discovery system to the task of image retrieval and show that the segmental SMT-based system achieves 72.2% recall@10 score in the phone-level. However, the drop in performance of all systems in the audio-level suggests a urgent need for better unsupervised phone-level and syllable-level representation for spoken language.

X. Acknowledgement

We would like to thank the 2017 JSALT team for their helpful discussions during and after the workshop, the 2017 ZRSC organizers for releasing the evaluation code and H. Kamper, O. J. Rašañen and others for releasing their codes for unsupervised word segmentation.

References

[1] S. Bharadwaj, M. Hasegawa-Johnson, J. Ajmera, O. Deshmukh, and A. Verma, "Sparse hidden Markov models for purer clusters," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[2] O. J. Rašañen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Interspeech*, 2015.

[3] H. Kamper, K. Livescu, and S. Goldwater. (2017) An embedded segmental k-means model for unsupervised segmentation and clustering of speech. [Online]. Available: https://arxiv.org/pdf/1703.08135.pdf

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

[5] C. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 40–49.

[6] L. Ondel, P. Godard, L. Besacier, E. Larsen, M. Hasegawa-Johnson, O. Scharenborg, E. Dupoux, L. Burget, F. Yvon, and S. Khudanpur, "Bayesian models for unit discovery on a very low resource language," in *Proc. ICASSP*, 2018.

[7] E. Dunbar, X. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," *CoRR*, vol. abs/1712.04313, 2017. [Online]. Available: http://arxiv.org/abs/1712.04313

[8] M. Johnson, T. Griffiths, and S. Goldwater, "Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models," in *Neural Information Processing Systems*, 2007.

[9] J. Trmal, M. Wiesner, V. Peddinti, X. Zhang, P. Ghahreman, Y. Wang, V. Manohar, H. Xu, D. Povey, and S. Khudanpur, "The Kaldi OpenKWS system: Improving low resource keyword search," in *Interspeech*, 2017.

[10] T. Alume, D. Karakos, W. Hartmann, R. Hsiao, L. Zhang, L. Nguyen, S. Tsakalidis, and R. Schwartz, "The 2016 BBN Georgian telephone speech keyword spotting system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[11] I. Medennikov, A. Romanenko, A. Prudnikov, V. Mendelev, Y. Khokhlov, M. Korenevsky, N. Tomashenko, and A. Zatvornitskiy, "Acoustic modeling in the STC keyword search system for OpenKWS 2016 evaluation," in *Interspeech*, 2017.

[12] Y. Khokhlov, N. Tomashenko, I. Medennikov, and A. Romanenko. (2017) Fast and accurate OOV decoder on high-level features. [Online]. Available: https://arxiv.org/pdf/1707.06195.pdf

[13] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Word segmentation through cross-lingual word-to-phoneme alignment," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2012, pp. 85–90.

[14] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2016, pp. 949–959. [Online]. Available: http://aclweb.org/anthology/N16-1109

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[16] P. Godard, M. Z. Boito, L. Ondel, A. Berard, A. Villavicencio, and L. Besacier, "Unsupervised word segmentation from speech with attention," in *Interspeech*, 2018.

[17] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: data, models and evaluation metrics," in *Journal of Artificial Intelligence Research*, 2010.

[18] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263 – 311, 1993.

[19] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Neural Information Processing Systems*, 2014.

[20] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the 2015 Conference on Computer Vision and Pattern Recognition, CVPR*, 2015.

[21] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," *Automatic Speech Recognition and Understanding*, 2015.

[22] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Neural Information Processing Systems*, 2016.

[23] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 613–622. [Online]. Available: https://www.aclweb.org/anthology/P17-1057

[24] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 27, pp. 89–98, 2019.

[25] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in 55^{th} *Annual Meeting of the Association for Computational Linguistics*, 2017.

[26] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 24, pp. 669–679, 2016.

[27] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[28] D. Harwath and J. Glass, "Towards visually grounded sub-word speech unit discovery," in *International Conference on Acoustics, Speech and Signal Processing*, 2019.

[29] L. Wang and M. Hasegawa-Johnson, "Multimodal word discovery and retrieval with phone sequence and image concepts," in *Interspeech*, 2019.

[30] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[31] S. Roweis, "Em algorithms for pca and sensible pca," California Institute of Technology, Tech. Rep., 1997.

[32] R. Fer *et al.*, "Multilingually trained bottleneck features in spoken language recognition," in *Computer Speech and Language*, 2017.

[33] P. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, 1992.

[34] J. K. Bock, D. E. Irwin, D. J. Davidson, and W. J. M. Levelt, "Minding the clock," *Journal of Memory and Language*, vol. 48, pp. 653–685, 2003.

[35] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," International Computer Science Institute, Tech. Rep., 1998.

[36] S. Vogel, H. Ney, and C. Tillman, "HMM-based word alignment in statistical translation," in *COLING '96 Proceedings of the 16th Conference on Computational Linguistics*, 1996.

[37] F. J. Och, , and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, 2003.

[38] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, 2014.

[39] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *IJCV*, vol. 123, no. 1, 2017.

[40] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. [Online]. Available: http://doi.acm.org/10.1145/219717.219748

[41] G. Adda *et al.*, "Breaking the unwritten language barrier: The bulb project," in *Procedia Computer Science*, vol. 81, 2016, pp. 8–14.