information

MDPI

Article

Subunits Inference and Lexicon Development Based on Pairwise Comparison of Utterances and Signs

Sandrine Tornay [1,2,*] **and Mathew Magimai.-Doss** [1]

[1] Idiap Research Institute, 1920 Martigny, Switzerland; mathew@idiap.ch
[2] Ecole polytechnique fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
* Correspondence: sandrine.tornay@idiap.ch; Tel.: +41-27-721-77-11

Received: 22 July 2019; Accepted: 24 September 2019; Published: 26 September 2019

Abstract: Communication languages convey information through the use of a set of symbols or units. Typically, this unit is word. When developing language technologies, as words in a language do not have the same prior probability, there may not be sufficient training data for each word to model. Furthermore, the training data may not cover all possible words in the language. Due to these data sparsity and word unit coverage issues, language technologies employ modeling of subword units or subunits, which are based on prior linguistic knowledge. For instance, development of speech technologies such as automatic speech recognition system presume that there exists a phonetic dictionary or at least a writing system for the target language. Such knowledge is not available for all languages in the world. In that direction, this article develops a hidden Markov model-based abstract methodology to extract subword units given only pairwise comparison between utterances (or realizations of words in the mode of communication), i.e., whether two utterances correspond to the same word or not. We validate the proposed methodology through investigations on spoken language and sign language. In the case of spoken language, we demonstrate that the proposed methodology can lead up to discovery of phone set and development of phonetic dictionary. In the case of sign language, we demonstrate how hand movement information can be effectively modeled for sign language processing and synthesized back to gain insight about the derived subunits.

Keywords: subword units; phone set; pronunciation lexicon; hidden Markov model; under-resourced; speech processing; sign language processing

1. Introduction

Communication is a vital part of human life. Speech is the most common mode of communication in the hearing community, while the preferred mode of communication in the deaf community is sign language to communicate. (Briefly, sign language is a visual mode of communication, where the information is conveyed through multiple visual channels such as hand gestures (hand shape, location, position and movement), facial expressions, body postures, and lip movements [1]. Similar to other natural languages, sign languages are natural languages with their own vocabulary and grammar, and they are different from the spoken languages. For example, British sign language is not a signed form of British English [1].) In terms of technology, spoken language technologies such as automatic speech recognition (ASR) systems and text-to-speech synthesis (TTS) systems have evolved over several years, and have reached a stage where they can be deployed. In contrast, although the methods have been borrowed from spoken language technology, sign language technology is still emerging [2]. One potential reason for that is that it has been well understood through linguistic studies that the time structure of spoken word units can be represented and modeled as a sequence of subword units such as phonemes/phones and syllables [3]. Such linguistically motivated subword units help in handing data scarcity issues when training statistical models and handle words that are unseen during training.

In the case of sign language, however, such a knowledge is still emerging. More precisely, it is yet to be understood well how the different channels (hand movement, hand shape, facial expression, etc.) in the visual signal together is perceived as a sequence of subword units [4,5], similar to how acoustic signal resulting from the movement of vocal folds and articulators (jaw, lips, and tongue) is perceived as sequence of phones or syllables in spoken language. This article focuses on subword units or subunits aspect. We use the terms subword units and subunits interchangeably in this article. Although the ensuing discussions in the remainder of this section are subword units related resource constraint challenges in speech processing, we show below that the research problems along those lines are relevant for sign language processing.

State-of-the-art methods for development of ASR systems and TTS systems presume that the target language has a written form and there exists a phonetic lexicon that transcribes the written form into sequence of phonemes/phones. Given the written form of words, a phonetic lexicon can be developed with the help of linguistic expertise or knowledge of the target language [3,6]. As a first step, a human expert manually transcribes each word into a phoneme sequence by observing the grapheme sequence (i.e., orthographic transcription). Once a base lexicon is available, a rule-based approach [7,8] or a learning-based approach (e.g., grapheme-to-phoneme conversion [9–11]) can be adopted to augment the lexicon with new words and pronunciation variants.

In the world, there are approximately 6900 languages and only about 5–10% of them employ a writing system [3]. Furthermore, not all of the languages that have a writing system may have a well developed phonetic dictionary. Studying these languages manually to acquire linguistic knowledge and phonetic dictionary from the acoustic data is a highly challenging and non-trivial task. Availability of computational methods can immensely help both the linguistic research community as well as the speech technology research community. One potential venue for that is the area of zero-resource speech processing (https://zerospeech.com/2015/index.html), which originally started with the problem of unsupervised speech pattern discovery [12,13], was then extended to automatic subword units discovery [14] and spoken term discovery [15], and more recently cast as a problem of automatic language acquisition by machines [16,17], taking inspirations from how infants and children acquire spoken language at the very early stages of life.

Instead of a completely unsupervised approach, yet another approach could be addressing a somewhat simplified question with light supervision: *Given only a set of utterances and the knowledge that any two pair of utterances correspond to the same word or not, how can the phone set inventory and a lexicon be automatically inferred*? Irrespective of whether the target language is known or not or whether it has a written form or not, this question can be posed. Furthermore, linguish notions such as minimal pairs are built on pairwise comparison. We can pose subword unit extraction in a similar manner. Availability of such data with pairwise comparisons can very well be envisaged in field linguistics, for instance collecting of speech utterances of day today life objects/entities (e.g., food, cloth, and numbers) possibly without the necessity to speak the unknown spoken language by showing them. In addition, if only acoustic data of an unknown language are available, such form of light supervision, i.e., whether two utterances correspond to the same word or not, could be obtainable from people with some speech expertise through listening tests. Furthermore, such a question can be posed in the above-mentioned zero resource speech processing framework after unsupervised spoken term for phone set or automatic subword unit inventory discovery and pronunciation model extraction. The same question could be posed in the case of sign languages to derive subunits and model signs as a sequence of subunits.

In this paper, we develop a hidden Markov models (HMM) based abstract framework that addresses the above posed scientific question with light supervision for linguistic resource development for speech processing and sign language processing, by building upon the inherent ability of HMMs to segment time series into stationary segments and recent works on resource-constrained speech processing using auxiliary multilingual resources. We validate the proposed framework through a spoken language study and a sign language study. Specifically, in the spoken language study, we demonstrate that the framework can lead up to phone set discovery and pronunciation lexicon

development. In the sign language study, a recognition-synthesis framework for hand movement subunits extraction and analysis is developed.

The remainder of the paper is organized as follows. Section 2 provides a brief survey of relevant literature. Section 3 presents the proposed framework. Sections 4 and 5 present the investigations on spoken language and sign language, respectively. Finally, in Section 6, we discuss the salient findings and conclude.

2. Relevant Literature

In the field of speech processing, the first methods for automatic subword units extraction and lexicon development emerged from the quest for finding alternatives to linguistically motivated subword units phones [18,19]. These methods typically involved [20,21]: (a) segmentation of speech utterances based on an acoustic similarity measure; (b) clustering of the segments using methods such as k-means to a *pre-set* number of subword units/clusters; and (c) finding pronunciations for each word from the occurrences of subword units in the training data followed by another clustering step to select representative pronunciations. These methods were then furthered in the context of pronunciation variation modeling [22,23]. Later, in the context of under-resourced speech processing, approaches have emerged where word level transcription of speech signal together with transcription of words in terms of graphemes have been exploited to derive automatic subword units and develop automatic subword units based lexicon [24–27], with an assumption that orthography of words and pronunciation of words in terms of phonemes are related. In other words, there exists a relationship between graphemes and phonemes. More recently, as mentioned above, subword units extraction in a completely unsupervised manner is being addressed for development of speech recognition and speech synthesis systems without prior linguistic knowledge from untranscribed speech [16,17].

In the field of sign language processing, there are subunit extraction approaches that assume that the produced signs have been annotated. For instance, Pitsikalis et al. developed an approach to convert HamNoSys (the Hamburg Notation System for Sign Languages—an alphabetic system describing the production of signs through the initial posture (describing nonmanual features, handshape, hand orientation and location) plus the actions changing this posture in sequence or in parallel [28]) representation into data-driven subunits [29]. Cooper et al. used hand labeled data and compared three types of subunits: appearance-based, 2D tracking-based and 3D tracking-based [30]. Koller et al. used gloss annotations and gloss time boundaries to generate sequences of subunits using HMM-based modeling and expectation-maximization algorithm [31]. (Gloss is a written form in sign language to provide semantic labels to sign. Usually, glosses take on the base form of a word in the most closely corresponding spoken language and are written in all upper case. For instance, in Swiss German Sign Language, the gloss "GESCHWISTER" represents "siblings") There are also subunit extraction approaches that do presume annotations or labeled data to be available. For instance, Bauer and Kraiss proposed an approach where first k-means is used to cluster the data obtained through hand gloves and then a sequence model is determined for each sign based on the clustered segments [32]. Han et al. used hand motion speed and trajectory information obtained through image processing to locate subunit boundaries and then adopting a temporal clustering by Dynamic Time Warping (DTW) approach to merge similar subunits [33]. Fang et al. segmented signs using HMM, in which each state represents one segment, and then they used a temporal clustering algorithm based on modified k-means algorithm, where DTW is employed for distance computation [34]. Theodorakis et al. built upon Fang et al.'s segmentation method and developed an HMM-based hierarchical clustering approach to extract subunits [35]. Sako and Kitamura employed a multi-stream HMM to segment signs and then applied a tree-based clustering algorithm to extract subunits [36].

In both speech processing and sign language processing, subword units or subunits extraction from the signal has been approached as a time series problem by employing segmentation and clustering techniques. Other than modality level difference (speech versus visual), the difference largely lies in the choice of the specific techniques employed for segmentation and clustering and the

order in which they are carried out. Thus, rather than approaching them as two different problems, subunits extraction for speech processing and sign language processing could be addressed through a common framework. Such a common framework could aid in making linguistic and technology advances in both fields. Besides lack of a common framework, there are a few other limitations in using the existing approaches as it is to address the light supervision problem posed in Section 1:

(i) Most of the methods for subword units extraction in speech processing employ or presume some form of prior knowledge, e.g., existence of writing system, availability of transcription, and linguistic knowledge to preset number of clusters. Zero resource speech processing, although alleviates the necessity of such prior knowledge, is still an emerging field. As stated in [16], the main focus is on "finding speech features that emphasize linguistically relevant properties of the speech signal (phoneme structure) and de-emphasize the linguistically irrelevant ones".

(ii) Unlike speech processing, in sign language processing, the acquisition methods have evolved such as from the use of hand gloves to camera systems. In addition, most of the subunits extraction investigations have been carried out in a signer dependent manner.

3. Proposed Approach

This section presents a methodology to address the light supervision computational linguistic problem posed in Section 1. More precisely, given only a set of utterances/sign productions for a spoken/sign language and the pairwise comparison between the utterances/sign productions, i.e., whether any two pair of utterances/sign productions in the set correspond to the same word (word can be regarded as an abstract unit of language, which can be realized in different ways to communicate [37], for instance written word, spoken word or sign production) or not, the goal is to automatically infer the subunits set and the associated lexicon that transcribes each word as a sequence of subunits. In the proposed methodology, given such pairwise comparison data, the following steps are performed:

Step 1: A sequence of feature vectors is extracted for each utterance or sign. In the case of speech signal, the feature vectors are short-term cepstral features, which tend to model information related to vocal tract system. In the case of signs, the feature vectors for hand movement are based on the 3D skeletal information.

Step 2: Given the sequence of feature vectors for each utterance or sign, a HMM is obtained for each unknown word or sign in the set. This step exploits the idea that HMMs inherently segment a time series into stationary segments and speech/sign recognition can be performed with word level HMMs.

Step 3: The states are clustered into subunits by pairwise comparison and a sequence model in terms of clustered subunits is obtained for each unknown word or sign.

Step 4: In the case of spoken language, phone set and pronunciation model for the unknown words are inferred by learning a probabilistic subunit-to-phone relationship exploiting auxiliary speech data with linguistic resources. As noted above, unlike speech processing, prior knowledge about how to model sign as a sequence of subunits is still emerging. Thus, in the context of sign language, visualization based on HMM-based synthesis is employed.

In the remainder of the section, we explain further Steps 2 and 3 in Section 3.1 and Step 4 in Section 3.2.

3.1. Automatic Subword Unit Based Lexicon Development

As illustrated in Figure 1, Steps 2 and 3 can be grouped together and seen as a step of deriving automatic subword units and development of an automatic subword units based lexicon. More precisely, in Step 2, word-level or sign-level HMMs for each word or sign are determined. This is done by modeling each state by a single Gaussian distribution with diagonal covariance and finding the

number of states, n, such that the recognition accuracy saturates on the training and the development data. Note that this process yields the same number of states n for all the unknown words or signs.

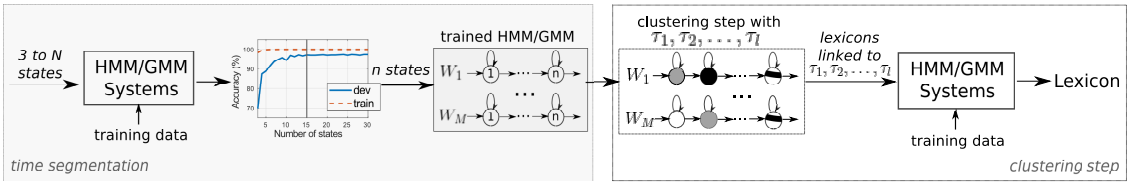

Figure 1. Illustration of the subword unit-based lexicon generation.

Given all the single Gaussians of the words or signs HMM states, Step 3 clusters them through a measure of discrimination. More precisely, this is done by computing, between each pair of Gaussian distributions, the Bhattacharyya distance [38]:

$$Bhatt(\mathcal{N}_1, \mathcal{N}_2) = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\ln\left(\frac{\det \boldsymbol{\Sigma}}{\sqrt{\det \boldsymbol{\Sigma}_1 \det \boldsymbol{\Sigma}_2}}\right), \quad (1)$$

where $\mathcal{N}_1 := \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2 := \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ are two Gaussian distributions and $\boldsymbol{\Sigma} := \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}$. The level of similarity between two HMM states is defined by a threshold τ, i.e., two HMM states are similar if the Bhattacharyya distance between the Gaussian distributions corresponding to the two states is below the threshold τ. The intuitive explanation is that two segments are modeling similar information or subword unit if the probability density functions (pdfs) of those states are similar. This clustering step yields a set of automatic subword units and an automatic subword unit based lexicon based on τ. The hyper-parameter τ is determined in a cross-validation manner as follows:

1. First, multiple automatic subword units based lexicons corresponding to different values of τ are obtained.
2. A recognition system is then trained on the training data based on each of those lexicons.
3. The lexicon that yields best recognition accuracy on the development data is chosen.

Selection of τ in this manner ensures that minimal discrimination between words or signs are maintained after the clustering step.

3.2. Linking Derived Subwords Units to Linguistic Knowledge

In Step 4, the goal is to establish a link to linguistic knowledge to ascertain the identity of the automatic subword units. In the case of spoken language, this can be done by learning a probabilistic relationship between the derived subword units and phones through acoustic signal. More precisely, as illustrated in Figure 2, this involves the following:

1. A phone posterior probability estimator on auxiliary data or languages that have well developed phonetic resources is trained.
2. A Kullback–Leibler divergence based HMM [39,40] (KL-HMM) with the phone posterior probability is trained as feature observations and the states being represented by the derived automatic subword units. Each state of the KL-HMM is parameterized by a categorical distribution of the same dimension as phone probability feature vector, which capture a probabilistic relationship between the automatic subword units and the phones. For a brief introduction about KL-HMM, the reader is referred to Appendix A. (The present paper builds upon different capabilities of KL-HMM: (a) modeling different subword units [41,42]; (b) handling resource constraints by exploiting multilingual or auxiliary resources [43]; and (c) modeling multichannel visual information in sign language [44]. For the sake of brevity, we do not go into very details of these works.)
3. Phone-based pronunciation is inferred by using the trained KL-HMM as a generative model and decoding the resulting sequence of probability through an ergodic HMM of phones.

It is worth mentioning that the proposed approach of inferring phonetic identities of the automatic subword units, and consequently a phonetic lexicon is inspired from the approach of acoustic data-driven grapheme-to-phoneme conversion using KL-HMM [42].

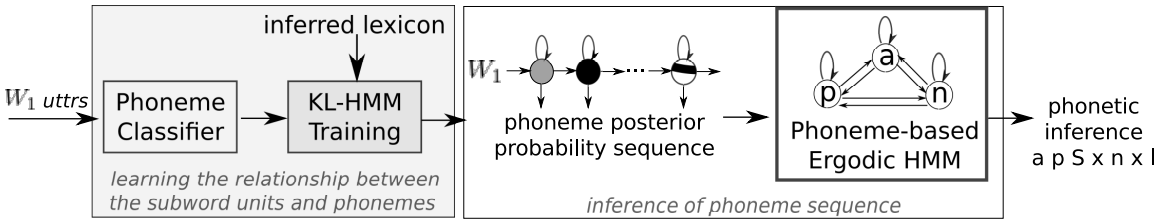

Figure 2. Illustration of the phoneme inference according to the derived subword units.

In the case of sign language, the subunits based lexicon represents the movement information for each observed sign "in-parts". It is not obvious to what prior linguistic knowledge those subunits could be linked. Even the HamNoSys annotation [28], which is used to transcribe signs, transcribes the whole movement information, not the movement information in-parts. Thus, we develop an alternate method, where the trained HMMs are used as a generative model to synthesize movement information in the 3D feature observation space by applying a linear-quadratic regulator (LQR) [45,46]. The synthesized movements for the signs can then be visually compared to the actual movements produced by the signers, which subsequently could be linked to HamNoSys should such annotation be available. Figure 3 illustrates the proposed approach to synthesize movement information of signs based on the derived subunits, starting from KL-HMM. As illustrated in the figure, given a sequence of Gaussian distributions, LQR finds the minimal path which passes through the sequence of the Gaussian distribution linked to each subunit that model a particular sign. The main question here being: Are the derived subunits able to synthesize the hand movement of signs such that it corresponds well with the human sign production? In that respect, a parallel between our approach and zero resource challenge 2019 (https://zerospeech.com/2019/) can be drawn, where a speech synthesis system based on automatic subword units derived in unsupervised manner is developed without using text transcription or phonetic transcription.

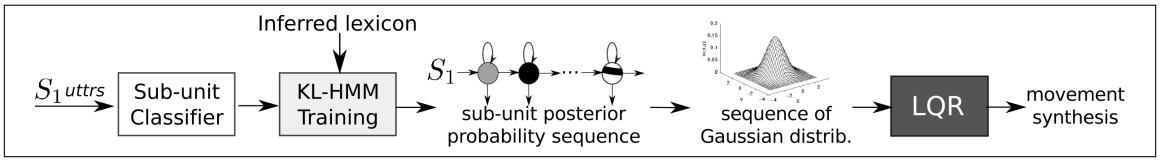

Figure 3. Illustration of the synthesis of the hand movement according to the derived subunits.

4. Spoken Language Study

In this section, we validate the proposed approach on a spoken language through ASR system level studies, as ASR relies on discrimination between words, and pronunciation lexicon level studies. To validate the proposed approach up to pronunciation level, we needed a language that has well developed linguistic resources as well as enough repetitions or acoustic realizations for each word from different speakers. Under-resourced languages typically lack such resources. Thus, we simulated the scenario through PhoneBook database that contains uncommon English words and proper names (e.g., Witherington and Gargantuan) produced by several speakers.

4.1. Experimental Setup

We used a part of the PhoneBook database for the study. We used 39-dimensional perceptual linear prediction (PLP) [47] cepstral features ($c_0 - c_{12} + \Delta + \Delta\Delta$) extracted with a window size of 25 ms and with a window shift of 10 ms as the feature vectors.

4.1.1. Database

PhoneBook is a speaker-independent phonetically-rich isolated-word telephone-speech English database [48]. PhoneBook consists of more than 92,000 utterances and almost 8000 different words, with an average of 11 different speakers/word. The database has been split into 106 word lists, each composed of around 75 words. Furthermore, the set of speakers is different for each word list. The word list contains uncommon English words and proper names (e.g., Witherington and Gargantuan). For our investigation, we used the small size (75 words) vocabulary setup; more precisely the *ad* word list that we separated into training, development and test sets as follows:

We selected speaker *m0k* who has uttered 74 words out of the 75 words as the development set. The development set was used for determining the number of HMM states per word in Step 2 and the clustering threshold τ in Step 3. With the ten remaining speakers, we performed a speaker independent experiment, where a leave-one-speaker-out protocol was applied. Thus, ten experiments were conducted, where, in each experiment, the data of one speaker were used for testing and the data of the remaining speakers were used for automatic subword units inference and for training ASR system. For each of the experiments, the average number of utterances for training, development and testing were 621, 74 and 69, respectively.

For lexical level validation studies, as part of Step 4, we used 21 word lists: *aa*, *ah*, *am*, *aq*, *at*, *ba*, *bh*, *bm*, *bq*, *bt*, *ca*, *ch*, *cm*, *cq*, *ct*, *da*, *dh*, *dm*, *dq*, *dt*, and *ea* to train phone-based classifier. This word list was originally defined in a study on task-independent speaker-independent speech recognition [49]: task-independent because the words in each word list are different and speaker-independent as the speakers in each word list are different. For example, words and speakers in word list *ad* are entirely different from the words and speakers in the 21 word lists. As done in [49], we used 42 context-independent phones (including silence) from the PhoneBook dictionary.

We also conducted a study where the phone posterior probability estimator was trained with multilingual data without English. For that, we used the Swiss French, Swiss German, Italian and Spanish part of the SpeechDat(II) corpus (http://www.speechdat.org/SpeechDat.html). Each language has about 12 h of speech. The lexicons are based on SAMPA phone set (https://www.phon.ucl.ac.uk/home/sampa/). We created a multilingual phone set by merging the phone sets across the four languages. This resulted in 104 context-independent phones including silence. It is worth mentioning that 35 phones out of the 104 phones are common to the English SAMPA phone set.

4.1.2. Systems

We built HMM with Gaussian mixture models (HMM/GMM) [50] and hybrid HMM/ANN (Artificial Neural Network) [51] systems to evaluate the automatic subword units based lexicon at ASR level. We built KL-HMM systems for lexical level validation. In each case, we built two systems: (a) using word-level HMM states obtained in Step 2 as subword units, referred to as word level system; and (b) using the clustered HMM states in Step 3 as subword units, referred to as clustered subword units based system. The motivation behind building word level system is that Step 2 obtains a word level HMM with fixed number of states n through discrimination, similar to in Step 3, so the states of the word level HMMs can also regarded as subword units without being clustered. Such a comparison would help us to determine whether the clustering step is indeed yielding meaningful subword units or not. The HMMs were trained and tested with the HTK toolkit [52]. The KL-HMM system studies were conducted using an in-house modified version of HTK. The neural networks, more precisely multilayer perceptrons (MLPs), for hybrid HMM/ANN and KL-HMM systems were trained using the Quicknet software [53].

HMM/GMM Systems: All the HMM/GMM systems are left-to-right HMMs using a single Gaussian distribution with diagonal covariance matrix as the emission distribution. In the case of the word level system, the number of states is chosen according to the model that saturates on the training and development data (Step 2). In the subword unit-based model, the clustering step was conducted with

the hyper-parameter, τ, in the range of 0.8–3.2 with a 0.2 step, each leading to a different lexicon. An HMM/GMM system was trained for each lexicon and the final one was chosen according to the maximum recognition accuracy on the development set (Step 3). The resulting word level system and clustered subword unit based system was tested on the test set. This process was repeated for each speaker-independent fold.

Hybrid HMM/ANN Systems: For building the hybrid HMM artificial neural network (ANN) systems, we first obtained the alignments in terms of the HMM states, respectively, from the word level and the clustered subword units-based HMM/GMM systems for each speaker-independent fold. We then trained MLPs classifying HMM states with output non-linearity of softmax and minimum cross-entropy error criterion. We used the 39-dimensional PLP cepstral features with four frames preceding context and four frames following context as the MLP input. In our experiments, we trained MLPs with different number of hidden units (600, 800, and 1000) and hidden layers (0, 1, 2, and 3). The number of hidden units and hidden layers as well as other hyper-parameters such as learning rate and the batch size were chosen according to the frame-level accuracy on the development set.

We estimated the scaled likelihoods in the hybrid HMM/ANN systems by dividing the posterior probabilities derived from MLPs with the prior probabilities of the classes estimated from relative frequencies in the training data. These scaled likelihoods were then used as emission probabilities for HMM states.

KL-HMM Systems: First, a single hidden layer MLP was trained to classify 42 context-independent phones, including silence. We used the 39-dimensional PLP cepstral features with four frames preceding context and four frames following context as the MLP input. The number of hidden nodes was 800. The KL-HMM parameters were then training by forward passing the training portion of the *ad* list data through the MLP and using resulting 42 dimensional phone posterior probability distribution per frame as the feature observations. We trained word level system and clustered subunits based system for each speaker-independent fold. After training the KL-HMM system, we tested the performance at ASR level on the test data. For lexical level validation, we generated the pronunciation of each word in terms of the 42 phones, as described in Step 4. For each word, we then computed the Levenshtein distance between the inferred pronunciation and the pronunciation given in the PhoneBook dictionary.

We trained a multilingual KL-HMM system for each speaker-independent fold where, we first trained a multilingual phone classifier on the SpeechDat(II) and then for each fold trained the KL-HMM parameters on the training portion of the *ad* list data, by forward passing it through the multilingual phone classifier and using the resulting multilingual phone posterior probabilities as feature observation.

4.2. Results and Analysis

In this section, we first present ASR level validation studies followed by lexical level validation studies. We then present as part of analysis: (i) impact of number of utterances on the proposed methodology on phone set and pronunciation model inference; and (ii) investigations using language independent multilingual data.

4.2.1. ASR Level Validation

First, the ASR study on the PhoneBook database was conducted to validate the assumption that the proposed approach derived discriminative subword units. As a baseline, we trained an HMM/GMM system with the lexicon based on the phonemes. Table 1 presents the average recognition accuracy (RA) over the ten-fold experiments of the clustered subword unit-based and word level systems. The average number of units corresponds to the number of HMM states used to train the model. In the case of the clustered subword unit-based system, a reduction of 28% of states/units can be observed.

Table 1. Clustered subword unit-based and word level systems RA on the PhoneBook database using PLP cepstral features with HMM/GMM and hybrid HMM/ANN systems.

	Clustered Subword Unit-Based System	**Word Level System**
HMM/GMM	94.1 ± 5.6	96.1 ± 4.0
Hybrid HMM/ANN	97.8 ± 2.0	98.3 ± 2.1
Average # units	*810*	*1125*
	Context-Independent System	**Context-Dependent System**
Baseline HMM/GMM	97.7 ± 2.5	96.1 ± 4.8
# phonemes	*39*	*114*

It can be observed that, in the case of HMM/GMM study, word-level system outperforms clustered subword units based system. However, in the case of hybrid HMM/ANN system, the performances are better than respective HMM/GMM system performance and are comparable. As a whole, the results indicate that the clustered subword units retain discrimination information across the words even with a reduction of around 28% of the number of HMM states.

Table 2 presents the average RA of the clustered subword unit-based and word level KL-HMM systems. As a baseline, we trained a KL-HMM system with the lexicon based on the phonemes. As can be seen, both systems yield comparable RAs, again indicating that clustered subword units retain discrimination across the words.

Table 2. KL-HMM-based subword unit- and word level -based systems results on the PhoneBook database using posterior distributions as features.

	Clustered Subword Unit-Based System	**Word Level System**
KL-HMM	99.0 ± 1.8	99.4 ± 1.2
	Context-Independent System	**Context-Dependent System**
Baseline KL-HMM	98.1 ± 3.4	99.7 ± 0.9

4.2.2. Lexical Level Validation

As explained in Section 3 (see Figure 2), we inferred the pronunciation of each word in the lexicon using the KL-HMM as a generative model, and decoded the resulting sequence of phone posterior probabilities for each word using a 42 phone fully connected ergodic HMM to get the pronunciation model. We compared the inferred pronunciations with the pronunciation provided in the PhoneBook dictionary, and computed Levenshtein distance (LEV) [54] and phone recognition rate (PRR). PRR is calculated as

$$1.0 - \frac{(\#insertion + \#deletion + \#substitution)}{N_{ref}}, \tag{2}$$

where # denotes "number of" and N_{ref} denotes the number of phones in the reference phonetic transcription. Table 3 presents the average LEV and PRR for pronunciations inferred by clustered subword unit based KL-HMM and word level KL-HMM. It can be observed that the inferred pronunciations are close to the original pronunciations in the manual pronunciation dictionary. Further analysis of the lexicon showed that the phonetic lexicon inferred using clustered subword unit based system cover 39 phones out of the 42 phones, while the manual dictionary for the words in *ad* list covers 38 phones. More precisely, with an exception of one extra phone, all the phones in the manual dictionary were inferred.

Table 3. Levenshtein distance (LEV) and phone recognition rate (PRR) results of the lexicon inferred from clustered subword unit-based KL-HMM system and word level KL-HMM system.

	Clustered Subword Unit-Based System	**Word Level System**
LEV $\pm$ std	1.9 $\pm$ 0.2	1.5 $\pm$ 0.1
PRR $\pm$ std	70.3 $\pm$ 2.6	76.4 $\pm$ 1.1

4.2.3. Further Analysis

Impact of number of utterances: In the experiments presented above, we had nine speakers utterances per word to derive subword units. In realistic under-resourced language scenario, it may not be possible to get so many speaker utterances per word. Thus, we studied the impact of number of speaker utterances on the proposed approach by developing two additional systems: (a) using only six speaker utterances per word (denoted as *six-utterances*); and (b) using only four speaker utterances per word (denoted as *four-utterances*) in a gender balanced manner. We compared the performances to the case where all the training utterances (denoted as *all-train-utterances*) were used. It is worth mentioning that, after deriving automatic subword unit lexicon, the HMM/GMM system was trained with all the utterances so that we can fairly compare the resulting lexicons. If the HMM/GMM systems were trained with four utterances or six utterances, separating the differences due to lexicon and data sparisty would have been a non-trivial task. Table 4 presents the average RA for HMM/GMM system. It can be observed that the amount of data used to infer automatic subword unit based pronunciation lexicon does not seems to affect the performance at ASR level. Interestingly, we can observe improvement with *six-utterances* based lexicon.

Table 4. Clustered subword unit-based and word level HMM/GMM systems results on the PhoneBook database depending on the three different setups used to infer the lexicon (*all-train-/six-/four-utterances* based lexicon) using PLP cepstral features.

	HMM/GMM-Based System		
	Lexicon	**Average RA $\pm$ std**	**Average # Units**
Clustered subword unit-based system	*all-train-utterances*	94.1 $\pm$ 5.6	*810 (−28%)*
	six-utterances	95.7 $\pm$ 4.5	*1365 (−9%)*
	four-utterances	95.4 $\pm$ 5.9	*1019 (−3%)*
Word level system	*all-train-utterances*	96.1 $\pm$ 4.0	*1125*
	six-utterances	96.3 $\pm$ 4.0	*1500*
	four-utterances	96.0 $\pm$ 5.5	*1050*

Table 5 presents the results with KL-HMM system at ASR level and lexical level. In this case, the automatic subword units based lexicon is derived using *all-training-*, *four-* or *six- utterances* and the KL-HMM is also trained on *all-train-*, *four-* or *six- utterances*, respectively. We can again observe that reduction in number of utterances does not affect Steps 2–4. Similar to the HMM/GMM study, at the ASR level, the number of samples does not impact the RA. At the LEV and PRR level, we can observe improvement with the *six-utterances* based lexicon.

Table 5. Clustered subword unit-based and word level KL-HMM systems recognition accuracy (RA), Levenshtein distance (LEV) and phone recognition rate (PRR) on the PhoneBook database depending on the three different setups used to infer the lexicon (*all-train-/six-/four-utterances* -based lexicon).

	Monolingual KL-HMM-Based System			
	Lexicon	**Average RA ± std**	**LEV ± std**	**PRR ± std**
Clustered subword unit-based system	*all-train-utterances*	99.0 ± 1.8	1.9 ± 0.2	70.3 ± 2.6
	six-utterances	99.3 ± 1.2	1.5 ± 0.1	76.2 ± 1.7
	four-utterances	99.0 ± 1.4	1.8 ± 0.1	71.5 ± 1.0
Word level system	*all-train-utterances*	99.4 ± 1.2	1.5 ± 0.1	76.4 ± 1.1
	six-utterances	99.3 ± 1.2	1.4 ± 0.0	77.5 ± 0.7
	four-utterances	99.1 ± 1.4	1.8 ± 0.1	72.1 ± 0.8

Multilingual study: In the above studies, the ASR level and lexical level studies were conducted in matched condition in terms of language. In other words, although the words and the speakers are not shared across word lists, the language is still English. Thus, we studied the possibility to use language-independent multilingual resources. For that, we performed ASR and pronunciation inference study using the multilingual KL-HMM system, where the 104-dimensional multilingual posterior probabilities estimated by MLP trained on Swiss French, Swiss German, Italian and Spanish portions of SpeechDat(II) were used as the feature observation to learn the relationship between the automatic subword units and the multilingual phones. Table 6 presents the ASR performance in terms of RA. For all cases, there is slight drop in performance when compared to the monolingual MLP. The trend remains that the same word level system and clustered subword units based system yield comparable systems. We also observe that reducing the number of utterances for derivation of automatic subword units and KL-HMM training does not impact the performance of the systems. This suggests that there exists a systematic relationship between the derived subword units and multilingual phones.

Table 6. Clustered subword unit-based and word level KL-HMM systems results on the PhoneBook database using multilingual phoneme classifier (without English).

	Multilingual KL-HMM-Based System	
	Lexicon	**Average RA ± std**
Clustered subword unit-based system	*all-train-utterances*	98.4 ± 1.5
	six-utterances	98.5 ± 1.6
	four-utterances	98.4 ± 2.5
Word level system	*all-train-utterances*	98.7 ± 2.1
	six-utterances	98.5 ± 1.6
	four-utterances	98.4 ± 2.5

We inferred the pronunciation based on the 104 multilingual SAMPA phone set. We found that *all-train-utterances* based lexicon covers 40 phones out of the 104 phones. Twenty-seven out of the inferred 40 phones belong to or shared to English SAMPA phone set, while 13 are borrowed from other languages. We were not able to carry out lexical level validation using LEV and PRR measures, as PhoneBook lexicon and SpeechDat(II) lexicon are based on two different Bets. It was not possible to map all the phones precisely, especially multilingual phones. Table 7 presents some examples of the phonetic inference in PhoneBook Bet for the monolingual case and SAMPA Bet for the multilingual case. We can observe that, unlike monolingual inference, the multilingual phone inference is somewhat noisy. This is potentially due to the mismatch in the database conditions. Compensating such differences and the Bet differences to further investigate the use of multilingual resources is part of our future study.

Table 7. Examples of phonetics inference according to the monolingual KL-HMM and multilingual KL-HMM.

Word	True	Monolingual-Based Inferrence	Multilingual-Based Inferrence
yarns	y a r n z	y a r n z	j o n
speechwriter	s p i C r Y t X	s p i C r Y t X	s p i tS u a OY e l
infrequently	I n f r i k w x n t l i	I n f r i k w x t l i	i e n f w i k u e
oops	u p s	w u p t s	n u
quail	k w e l	k w e l	u w e i o
bonbon	b a n b a n	b @ a n b a x n	o n b o n

5. Sign Language Study

In this section, we validate the proposed approach on sign language. We limited the investigations to hand movement and hand shape information, as extraction of non-manual features such as mouthing or facial expressions is still an open research problem [2,55]. In particular, we focused on deriving hand movement subunits and study their modeling with and without hand shape information.

5.1. Experimental Setup

As part of validation, we conducted *signer-independent* automatic sign language recognition (SLR) and synthesis of hand movement information from the resulting models. The study was conducted on the SMILE Swiss German Sign Language database using hand movement information. Speech technologies based approaches benefit from the idea that subword units such as phones can be shared across languages [43]. In the spoken language part, we used that capability in the study using auxiliary multilingual resources. We took inspiration from that to investigate: whether the derived hand movement subunits exhibit similar desirable characteristics? For that, we used Turkish sign language data to perform cross-lingual sign language processing experiments.

5.1.1. Database

The large-scale SMILE Swiss German Sign Language Dataset [56] was created in the context of developing an assessment system for lexical signs of Swiss German Sign Language (DSGS) (Deutschschweizerische Gebärdensprache). It contains 11 adult signers and 19 adult L2 learners which produced 100 isolated signs of a DSGS vocabulary production test. Each sign was performed three times and only the second pass was manually annotated. The SMILE dataset was collected with the Microsoft Kinect v2 sensor and the high speed and high resolution GoPro video cameras. The SMILE dataset provides the color videos, depth maps, user masks and 3D pose information obtained from the Kinect, and the body pose, facial landmarks, and hand pose information extracted using the deep-learning-based key point detection library OpenPose.

As mentioned, only the second pass was annotated through six categories that evaluate the acceptability of a sign according to linguistic criteria; more precision can be found on the "Category of sign produced" annotation of the SMILE transcription/annotation scheme presented in [56]. In our experimental studies, we only used the second pass data that were annotated as Category 1 or 2, i.e., acceptable signs with the same or slightly the same form. We did not make any difference between the L1 and L2 signers in our experiments. To ensure enough samples are available for each sign (minimum five samples/sign), 94 signs were selected out of the 100. The resulting data for the 94 signs were partitioned in a signer-independent manner into 1263 training set samples from 17 signers, 249 development set samples from 3 signers and 704 test set samples from 10 signers.

To investigate the ability to share the derived subunits across sign languages, we used Turkish sign language (TSL) HospiSign database. HospiSign is a subset of 33 TSL phrase classes related to the health domain of the continuous BosphorusSign database [57]. The HospiSign subset includes six signers,

with each sign being repeated approximately six times by each signer. The database is available upon request from the authors (https://www.cmpe.boun.edu.tr/pilab/BosphorusSign/home_en.html). The database has been recorded with a Kinect camera. We used the skeletal joint coordinates that are provided in the database for feature extraction.

5.1.2. Feature Extraction

Two types of feature were used to model the hand movement: (i) the 3D skeleton position of both hands according to three different coordinate systems (based on the head center, the hand corresponding shoulder center and hip center) normalized by the head width; and (ii) the corresponding delta velocities. More precisely, for each time frame, we first normalized position features of the left and right hand, $\mathbf{p_t}$, by the width of the head. Then, three types of 3D coordinate of the hands were recalculated depending on three coordinate systems. The first one takes the head as the center; the second one uses the right shoulder as the center for the right hand, and uses the left shoulder as the center for the left hand; and the third one takes the right hip as the center for the right hand, and takes the left hip as the center for the left hand. Therefore, depending on the center $\mathbf{C}$, the position feature $\mathbf{p_t}$ would be:

$$\mathbf{p}_t^{\mathbf{C}} = \frac{\mathbf{hand} - \mathbf{C}}{|\mathbf{head} - \mathbf{neck}|/4'} \tag{3}$$

where $\mathbf{C} \in \{\mathbf{head}, \mathbf{shoulder}, \mathbf{hip}\}$; **hand**, **shoulder**, **hip** are vectors of $\mathbf{x}, \mathbf{y}, \mathbf{z}$ coordinates of, respectively, left and right hand, shoulder and hip; and **head** contains $\mathbf{x}, \mathbf{y}, \mathbf{z}$ coordinate of the head twice.

The velocity features, $\mathbf{v_t^C}$, are estimated for each coordinate system by computing the difference between the position features at time t and time $t-2$.

$$\mathbf{v_t^C} = \mathbf{p_t^C} - \mathbf{p_{t-2}^C}. \tag{4}$$

The resulting features are of size 36, 18 positions features ((3 left + 3 right hand position features) $\times$ 3 coordinate systems) and 18 velocity features.

5.1.3. Systems

As done in the case of spoken language study, we used HTK, Quicknet and in-house implementation of KL-HMM on HTK for developing the sign language recognition systems. Similar to spoken language investigations, we studied the *sign level* system based on the HMM states in Step 2 and clustered subunits-based system based on Step 3. Unlike spoken language study, there is no phonetic dictionary which we can infer. Thus, we conducted visualization studies by applying LQR-based hand movement information synthesis using the *pbdlib* library, developed by Pignat and Calinon in [45] in the context of robotics.

HMM/GMM Systems: All the HMM/GMM systems are left-to-right HMMs using one mixture Gaussian distribution with diagonal convariance matrix per state as the emission distribution. In Step 2, the number of states for the sign level HMM is chosen according to the saturation of the model on the SMILE training and development data. In Step 3, the clustering step was conducted with the hyper-parameter, τ, in the range of 0.3–1.3 with a 0.1 step, leading to a set of lexicon. An HMM/GMM system was trained for each lexicon and the one that yields the maximum recognition accuracy on the development set was chosen. Test set performances are reported on that lexicon.

Hybrid HMM/ANN Systems: For building the hybrid HMM artificial neural network (ANN) systems, we first obtained the alignments in terms of the HMM states using either the sign level or the clustered subunits-based HMM/GMM systems. We then trained MLPs classifying HMM states with output non-linearity of softmax and minimum cross-entropy error criterion. We used the 36-dimensional movement features with four frames preceding context and four frames following context as the MLP

input. In our experiments, we trained MLPs with different number of hidden units (600, 800, and 1000) and hidden layers (0, 1, 2, and 3). The number of hidden units and hidden layers as well as other hyper-parameters such as learning rate and the batch size were chosen according to the frame-level accuracy on the development set.

We estimated the scaled likelihoods in the hybrid HMM/ANN systems [51] by dividing the posterior probabilities derived from MLPs with the prior probabilities of the classes estimated from relative frequencies in the training data. These scaled likelihoods were then used as emission probabilities for HMM states during decoding.

KL-HMM Systems: We developed two different left-to-right KL-HMM systems:

1. Monolingual KL-HMM system: In this case, the hand movement subunits posterior probabilities estimated by the MLP of hybrid HMM/ANN system are used as feature observations. The KL-HMM states represent the hand movement subunits.
2. Cross-lingual KL-HMM system: In this case, the hand movement subunits are derived on TSL HospiSign database (Steps 2 and 3); an MLP is trained on the HospiSign data to estimate TSL subunits posterior probabilities; and the states model DSGS subunits and the parameters are trained by using the TSL subunits posterior probabilities estimated on the DSGS data as feature observations. In doing so, the KL-HMM learns a probabilistic relationship between DSGS subunits and TSL subunits, and allows us to examine language-independence of derived subunits. To compensate the difference in the coordinate system recording in between both databases, a skeleton alignment is applied before the feature extraction. To do so, all the signer skeletons of both databases are aligned at the neck joint with respect to a reference Hospisign signer skeleton and then scaled by the shoulder width.

Following a recent work [44], in the KL-HMM framework, we also studied combining subunits based hand movement information and hand shape information for sign language recognition. The hand shape features used in this paper are the output posterior distributions of the Deep Hand estimator developed by Koller et al. [58]. The estimator was trained on the one-million hands dataset [58]. The set of hand shape classes used to train the estimator consists of a transition shape and the 60 hand shapes (linguistically inspired) presented in https://www-i6.informatik.rwth-aachen.de/~koller/1miohands-data/. The resulting hand shape features are a 122-dimensional features containing the 61 dimensional posterior distribution for each hand.

5.2. Results and Analysis

We first present the monolingual study and then present the cross-lingual study.

5.2.1. Monolingual Study

Table 8 presents the sign language recognition accuracy (RA) depending clustered subunit-based system and sign level system on the SMILE database along with the average number of subunits in each case. It can be observed that the clustered subunits-based system with around 14% less HMM states performs comparable to sign level system for both HMM/GMM and hybrid HMM/ANN system. This indicates that the clustered subunits-based lexicon obtained in Step 3 maintains discrimination across signs. Furthermore, KL-HMM systems trained using subunits posterior probability as feature observation further improves the recognition accuracy.

Table 8. Hand movement clustered subunits-based and sign level HMM/GMM, hybrid HMM/ANN and KL-HMM systems performance in terms of recognition accuracy on the SMILE database.

	Clustered Subunit-Based System	Sign Level System
HMM/GMM	51.3	49.4
Hybrid HMM/ANN	51.6	53.0
KL-HMM	55.8	57.4
Average # subunits	*1945*	*2256*

To get an insight into how consistent are the derived subunits, we synthesized the 3D hand position movement with a LQR using the sequence of Gaussian distributions linked to the sequence of subunits of the sign (Step 4). The starting point for the hand movement synthesis is the starting point of a manual production of the sign. The duration for each state is the average number of time frames estimated by aligning the states in the sign to the training samples. Two signs are presented: TAXI, which is a well-recognized sign (100% recognition), and PAPIER, which is a poorly recognized sign (0% recognition). To facilitate the visualization, we depict in Figure 4 the (x, y) position of the dominant hand as well as for comparison three examples of the respective sign samples (soft lines), the z-axis, the depth of the sign production, being not relevant for these two particular signs.

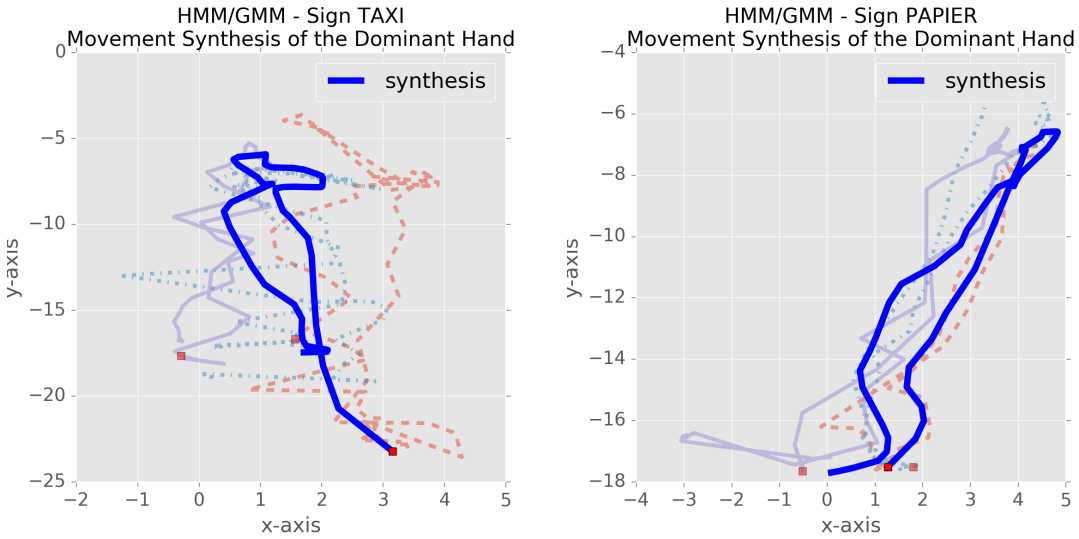

Figure 4. Movement synthesis of the dominant hand for the well-recognized sign, TAXI (**left**), and the poorly-recognized sign, PAPIER (**right**), using the Gaussian distribution sequence of the subunit-based HMM/GMM system, where the red square are the starting points.

As can be seen, the hand movement of signers vary a lot. (For the sake of clarity, we did not show all the signers production.) Nevertheless, in both cases, the synthesized movements follow similar direction and range of movement as the hand movement of the actual signers. This suggests that the subunits are modeling the relevant hand movement information.

The sequence of the Gaussian distributions can be also obtained according to the KL-HMM system. The advantage of the KL-HMM is its categorical distributions states, which allows computing a new Gaussian by using them as a weight on the subunit Gaussians. First, we selected significant values, i.e., categorical distributions greater than 0.005, and re-scaled them according to the total number of selected components, M. Then, the combined mean, μ_{comb}, and diagonal covariance, σ_{comb}, were computed as:

$$\mu_{comb} = \frac{1}{M} \sum_{m=1}^{M} \mu_m \quad , \qquad \sigma_{comb} = \sum_{m=1}^{M} w_m \cdot \sigma_m,$$

where w_m are the re-scaled categorical distributions and μ_m, σ_m the mean and the diagonal covariance of the corresponding Gaussian distribution. Figure 5 depicts the resulting movement synthesis of the well-recognized sign, TAXI (100% recognition), and WASCHEN, a poorly-recognized sign (0% recognition). When comparing the synthesized movement for sign TAXI across KL-HMM and HMM/GMM (Figure 4), the difference mainly appears at the end of the sign.

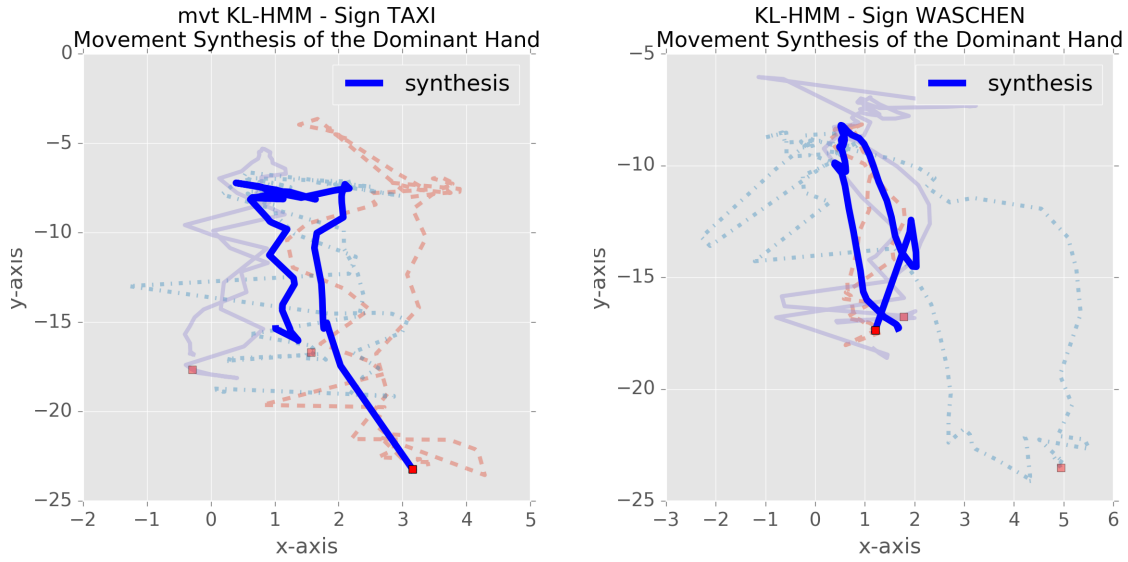

Figure 5. Movement synthesis of the dominant hand for the well-recognized sign, TAXI (**left**), and the poorly-recognized sign, WASCHEN (**right**), using the Gaussian distribution sequence computed from the subunit-based KL-HMM system, where the red square are the starting points.

As the synthesized movement of the dominant right hand of the poorly-recognized sign, PAPIER, is corresponding to the movements produced by the signers, we looked at the confusion matrix to understand the reason for the poor recognition accuracy and analyzed the hand movements. It was found that the hand movements of some signs are similar, as can be seen in Figure 6.

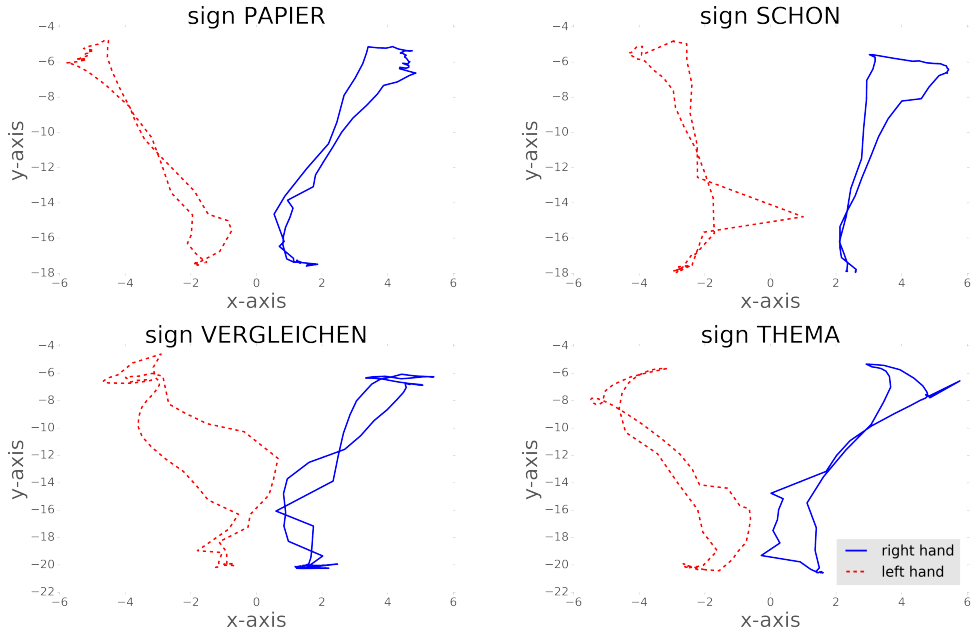

Figure 6. (x, y) movement of the right and left (dashed line) hands of the signs PAPIER, SCHON, VERGLEICHEN and THEMA, respectively.

Sign language convey information through multiple channels. A single channel (e.g., only hand movement) may not be sufficient to discriminate all the signs. The confusion in terms of hand movements can be handled by adding other channel such as hand shape to the KL-HMM system. Table 9 presents the sign language recognition of the KL-HMM system only with hand movement features (*hmvt*), only with hand shape features (*hshp*) and both (*hmvt+hshp*). As it can be seen, use of both hand movement information and hand shape information increases the recognition performance. Indeed, in the case of the sign PAPIER, integration of hand shape information increases the RA from 0% to 80%.

Table 9. Sign RA on the DSGS database of the KL-HMM system depending on the movement (*hmvt*) DSGS subunits posterior probabilities, the hand shape (*hshp*) features or both (*hmvt+hshp*) of them.

	hmvt **KL-HMM**	***hshp*** **KL-HMM**	***hmvt+hshp*** **KL-HMM**
Sign RA	55.8	38.2	74.3

A question that arises is: Does adding the hand shape information affect the model of the movement subunits? Thus, we synthesized hand movements using the hand movement categorical distributions estimated by *hmvt+hshp* KL-HMM. Figure 7 depicts the dominant hand movement synthesized by the *hmvt* KL-HMM system and of the *hmvt+hshp* KL-HMM system. As can be observed, the synthesized movements are slightly different, with the *hmvt+hshp* KL-HMM following better the hand movement in the manual productions of the sign.

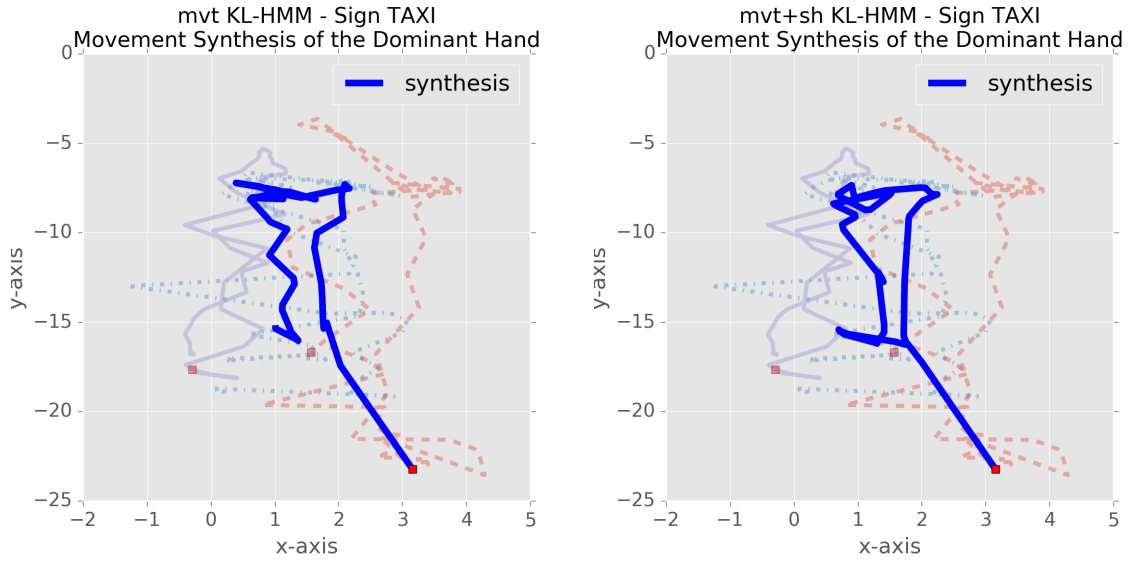

Figure 7. Movement synthesis of the dominant hand for the sign TAXI using the Gaussian distribution sequence computed from the *hmvt* KL-HMM system and the *hmvt+hshp* KL-HMM system. The red squares are the starting points.

5.2.2. Cross-Lingual Study

Table 10 presents the results of the cross-lingual study, where the subunits are derived on the TSL HospiSign and KL-HMM system is trained on DSGS database to recognize DSGS signs with DSGS subunits lexicon. It can be observed that the performance drops considerably when compared to the monolingual case. However, the performance obtained is beyond chance level. This suggests that there exists some degree of systematic relationship between the DSGS subunits and TSL subunits but it is not sufficient to recognize well the DSGS signs. The reason for that could be: (a) differences in the coverage of hand movements across the two databases; and (b) differences in recording settings.

In the case of HospiSign, the signs were performed in standing position, while, in the case of DSGS, the signs were performed in sitting position. Skeleton alignment may not have fully compensated for these differences. Investigating these differences and their impact is part of our future work.

Table 10. Sign language RA on the DSGS database of the KL-HMM system trained with TSL HospiSign subunits posterior probabilities in the multilingual case and DSGS subunits in the monolingual one.

	Cross-Lingual System	Monolingual System
Sign RA	41.5	55.8

Figure 8 depicts the synthesis of the movement based on the TSL HospiSign subunits of a well-recognized sign and poorly-recognized sign, THEMA and SPIELEN, respectively. In the case of sign THEMA, we can observe that the synthesized movement follow hand movement of the actual signers. In the case of sign SPIELEN, it is not the case. One of the reasons for that could be that TSL subunits may not be covering well all the DSGS movements.

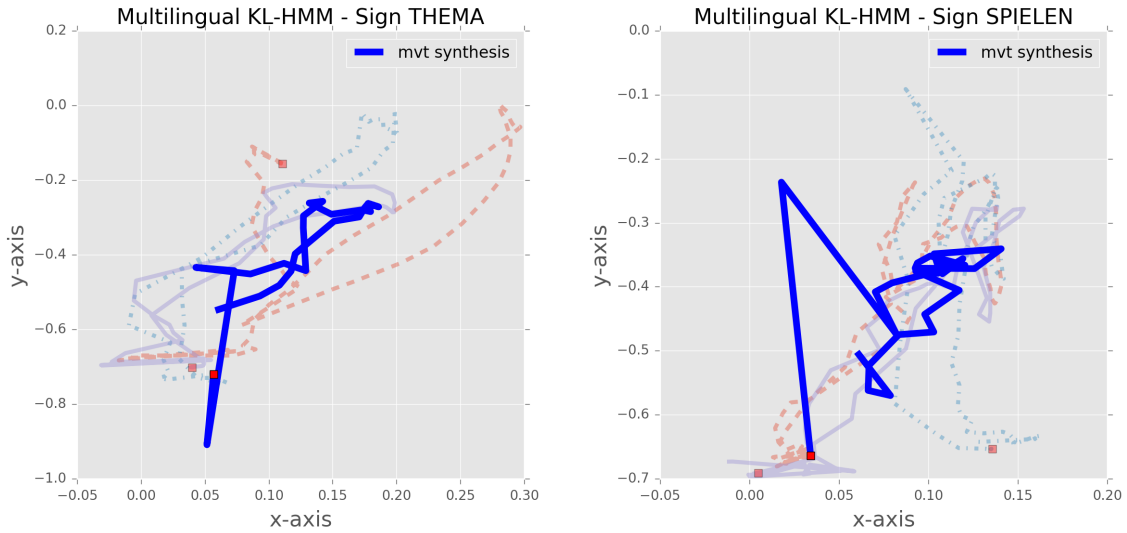

Figure 8. Movement synthesis of the dominant hand for the well-recognized sign, THEMA (**left**), and the poorly-recognized sign, SPIELEN (**right**), using the Gaussian distribution sequence computed from the KL-HMM system using the TSL HospiSign subunits. The red squares are the starting points.

Table 11 presents the performance of the cross-lingual system when hand shape information is also modeled. The performance of the system improves significantly. The gap between monolingual and cross-lingual system reduces from 13.3% absolute (55.8–41.5) to 8.2% absolute (74.3–66.1). This suggests that some shortcomings of cross-lingual hand movement information estimation is getting compensated by hand shape information. At the same time, it should be noted that hand shape alone yields performance of 38.2%. Thus, the improvement in performance is attributed to the complementary discrimination provided by both hand movement and hand shape, similar to the monolingual case. Together, these results indicate that the derived subunits could be shared. The ability to share across languages could potentially be improved by using multiple sign language data.

Table 11. Sign RA on the DSGS database of the cross-lingual KL-HMM system depending on the movement (*hmvt*) TSL HospiSign subunits posterior probabilities alone or added to the hand shape features (*hmvt+hshp*).

	hmvt **KL-HMM**	***hmvt+hshp*** **KL-HMM**
Sign RA	41.5	66.1

Impact of number of training samples: As mentioned above, sign languages are inherently under-resourced. Only few sign languages have a proper database. This is further compounded by the fact that sign languages are highly localized and may not be related to each other. For instance, although English is the native language for Brits and Americans, British sign language (BSL) and American sign language (ASL) are distinct languages [1,59]. In other words, while American English and British English are mutually intelligible, BSL and ASL are not. Similarly, German sign language (DGS) and DSGS are distinct sign languages. The cross-lingual approach is leading towards methods to share resources across sign languages. Thus, we investigated the impact of number of training samples per sign from the target sign language on the performance of cross-lingual sign language recognition system.

In the DSGS database, since we only used Category 1 and Category 2 data from SMILE dataset (see Section 5.1.1), the number of training samples varies; Figure 9 depicts the histogram of the number of samples per sign. To find the appropriate number of training samples per sign, we conducted a study using three different setups where in the first one ten samples per sign (referred as *ten-sample-signs*), in the second eight samples per sign (*eight-sample-signs*) and in the last six samples per sign (*six-sample-signs*) were used. To evaluate the three different setups, we derived hand movement subunits for each setup and built KL-HMM based sign language recognition systems. We also evaluated them with the KL-HMM system modeling both the hand movement and hand shape information.

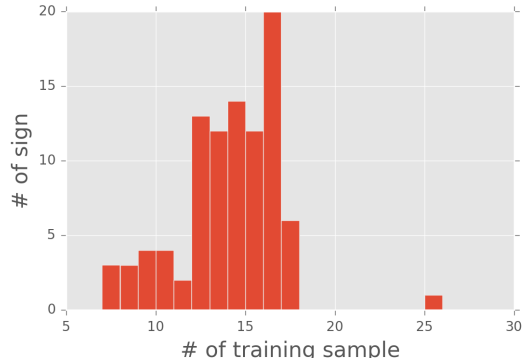

Figure 9. Histogram of the number of training samples per sign of the DSGS dataset.

Table 12 presents the RA of this study.

Table 12. Cross-lingual KL-HMM based *hmvt* and *hmvt+hshp*-based systems results on the DSGS database depending on the three different setups used to infer the lexicon (*ten-/eight-/six-sample-signs lexicon*).

	Lexicon	**Sign RA**
hmvt KL-HMM System	*ten-sample-signs*	35.9
	eight-sample-signs	33.7
	six-sample-signs	33.8
hmvt+hshp KL-HMM System	*ten-sample-signs*	62.6
	eight-sample-signs	60.1
	six-sample-signs	55.0

It can be observed that the RA decreases with decrease in number of samples per sign. As can be observed in Figure 10, the minimum number of samples seems to be around twelve samples, but the figure also shows that the number of samples needed depends on the sign. The different movement complexity of the signs and variations introduced by the signers can explain this difference. Furthermore, as we observed earlier, the differences in the coverage of hand movements across the two database can also explain why a low number of samples is not sufficient. Investigating further this aspect with multilingual sign language resources is part of our future work.

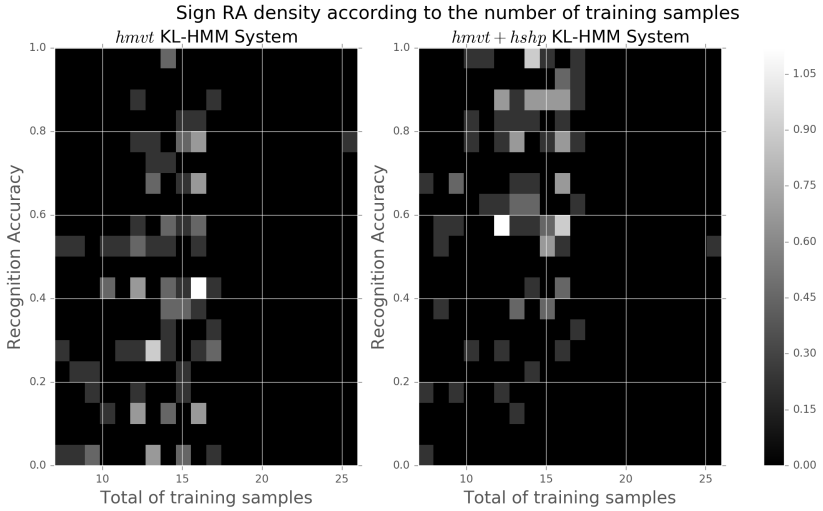

Figure 10. Sign RA density according to the number of training samples according to the KL-HMM system using the TSL movement subunit (*hmvt*) or using both with the hand shape information (*hmvt+hshp*).

6. Discussion and Conclusions

This paper addresses a computational linguistic paradigm, where, given a set of speech utterances or sign productions and the pairwise comparison between each pair of speech utterances or sign productions in the set on whether they correspond to the same word or sign or not, the goal is to derive subword units or subunits, and link them to available prior linguistic knowledge. Toward that goal, we present a methodology which first obtains a word level or sign level HMM for each lexical entity and clusters the HMM states by pairwise discrimination to obtain a reduced set of states or subunits and a corresponding subunits based lexicon. In the case of spoken language, we demonstrated that, for both the states of word level HMM and derived subword units (i.e., clustered states), a probabilistic relationship to phones can be learned by exploiting auxiliary resources to identify the phone set and obtain a phone-based pronunciation dictionary. In the case of sign language, due to lack of concrete linguistic knowledge to link hand movement subunits, we propose a visualization method where the hand movement information is synthesized to assess how well the subunits model the movement information relevant to signs. As a by-product, the proposed methodology also leads to methods to develop technologies in such a resource-constrained scenario.

In the experimental studies on spoken language as well as sign language, we observed that both use of the states of word level or sign level HMMs and use of the clustered HMM states as subword units lead to comparable recognition performances. This is highly encouraging and interesting. In the speech recognition literature, it has been observed that word level HMMs outperform subword units based modeling [18,19]. However, our investigations, in particular lexical level validation in the spoken language study, also point out that Step 3, i.e., clustering of HMM states, may not be necessary, as the states of word level HMMs can be linked to the prior knowledge to discover the phone set and

obtain phonetic lexicon. Thus, whether to cluster the word level HMM-states or not is open for further investigation.

In the context of sign language, Table 13 compares our work with a few closely related subunit extraction studies. As can be seen, the previous approaches have focused on processing images or motion information captured via gloves, while our approach focuses on modeling skeleton information, which can be easily and reliably obtained nowadays. In addition, most of these works have not investigated signer independence. Furthermore, none of these works have studied cross-lingual aspects. Finally, our framework allows synthesis of hand movement in 3D space to better understand the derived subunits using LQR. In doing so, our work establishes a link between modeling hand movement in sign language and in robotics.

Table 13. Comparison of our sign language hand movement subunit studies with existing studies.

Ref.	Features Based	Segment.	Clustering Algorithm	Recognition Study	Signer Indep. Study	Monolingual/ Cross-Lingual
Sako and Kitamura [36]	images processing	multi-stream HMM	tree based algorithm	✓	✓	Monolingual
Bauer and Kraiss [32]	gloves	HMM	*k*-means	✓	✗	Monolingual
Han et al. [33]	images processing	discontinuity detector	DTW	✓	✗	Monolingual
Fang et al. [34]	gloves	HMM	modified *k*-means	✗	✗	Monolingual
Theodorakis et al. [35]	images processing	HMM	HMM hierarchical clustering	✗	✗	Monolingual
Our approach	skeleton	HMM	pair-wise clustering with Bhatt. dist.	✓	✓	Mono- and cross-lingual

Our future work will consider the following directions:

1. Further ground the methodology linguistically by: (a) modeling articulatory features [60,61] instead of cepstral features in the case of spoken language; and (b) modeling both hand movement and hand shape in the case of sign language for deriving subunits. In both cases, it can be achieved by employing KL-HMM in Step 2 and Step 3. In the case of sign language, it will help in connecting to linguistic research that are trying to understand the high level units formed by hand movement and hand shape [4,5].
2. Extend the cross-lingual investigations on sign language to multilingual scenario by pooling resources from other sign languages, with the ultimate aim of handling resource constraints in sign language processing.
3. In the speech community, there is interest in modeling articulatory measurements obtained through electromagnetic articulography [62,63]. In the present work, the hand movement subunits were extracted by modeling skeletal information, i.e., measurements in 3D coordinate system. We will investigate whether such a method can be adopted to derive subunits from articulatory measurements, which in turn could be related well to the acoustic signal and phones.

Author Contributions: Conceptualization, M.M.-D. and S.T.; methodology, M.M.-D. and S.T.; software, S.T.; validation, S.T. and M.M.-D.; formal analysis, S.T.; investigation, S.T.; resources, S.T. and M.M.-D.; data curation, S.T. and M.M.-D.; writing—original draft preparation, S.T. and M.M.-D.; writing—review and editing, M.M.-D. and S.T.; visualization, S.T.; supervision, M.M.-D.; project administration, M.M.-D.; and funding acquisition, M.M.-D.

Funding: This research was funded by the Swiss National Science Foundation through the Sinergia project SMILE *(Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment)*, grant agreement CRSII2_160811.

Acknowledgments: We thank all the collaborators in the project SMILE for their valuable work. In particular, Necati Cihan Camgöz (CVSSP, University of Surrey) for providing us the hand shape estimator and the resulting hand shape posterior probability features and Marzieh Razavi (presently with Telepathy Labs, Zürich) with the implementation of KL-HMM system. We also thank Emmanuel Pignat (Robot Learning & Interaction group, Idiap

Research Institute) for providing us the code of the LQR to synthesize the hand movement coming from his *pbdlib* Python library.

Conflicts of Interest: The authors declare no conflict of interest. The funding agency had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
ASL	American Sign Language
ASR	Automatic Speech Recognition
BSL	British Sign Language
DGS	German Sign Language
DSGS	Swiss German Sign Language
DTW	Dynamic Time Warping
GMM	Gaussian Mixture Model
HamNoSys	Hamburg Notation System
HMM	Hidden Markov Model
HTK	Hidden Markov Model ToolKit
KL-HMM	Kullback–Leibler divergence-based HMM
LEV	Levenshtein distance
LQR	Linear-Quadratic Regulator
MLP	MultiLayer Perceptron
PLP	Perceptual linear prediction
PRR	Phone Recognition Rate
RA	Recognition Accuracy
SAMPA	Speech Assessment Methods Phonetic Alphabet
SLR	Sign Language Recognition
TSL	Turkish Sign Language
TTS	Text-to-speech Synthesis

Appendix A. Kullback–Leibler Divergence Based HMM

Figure A1 illustrates Kullback–Leibler divergence based HMM (KL-HMM) system in the context of speech processing. In this system, each state i of the HMM is parameterized by a reference multinomial or categorical distribution $\mathbf{y}_i = [y_i^1, \cdots, y_i^D]^T$, where D is number of subword units or subunits. In the illustration, the subword units are phonemes. The state transition probabilities are typically set to 0.5 for self transition and 0.5 to leave the sate.

Given an estimate of posterior feature observation $\mathbf{z}_t$ at time frame t,

$$\mathbf{z}_t = [z_t^1, \cdots, z_t^D]^T = [P(p_1|\mathbf{x}_t), \cdots, P(p_D|\mathbf{x}_t)]^T$$

the local score at each HMM state is estimated as Kullback–Leibler (KL) divergence between $\mathbf{y}_i$ and $\mathbf{z}_t$, i.e.,

$$KL(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log(\frac{y_i^d}{z_t^d}) \tag{A1}$$

where $\mathbf{x}_t$ is the acoustic feature (such as cepstral feature) at time frame t, $\mathbf{y}_i$ is the reference distribution, and $\mathbf{z}_t$ is the test distribution. We denote this local score as KL.

KL-divergence being an asymmetric measure, there are also other ways to estimate the local score,

1. Reverse KL-divergence (*RKL*):

$$RKL(\mathbf{z}_t, \mathbf{y}_i) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \tag{A2}$$

2. Symmetric KL-divergence (SKL):

$$SKL(\mathbf{y}_i, \mathbf{z}_t) = KL(\mathbf{y}_i, \mathbf{z}_t) + RKL(\mathbf{z}_t, \mathbf{y}_i) \tag{A3}$$

The parameters of the HMM states (i.e., multinomial distributions) are trained using Viterbi expectation maximization algorithm with a cost function based one of the local scores [39,40]. The decoding is performed using standard Viterbi decoder, where the state emission log-likelihood is replaced by the KL-divergence based local score.

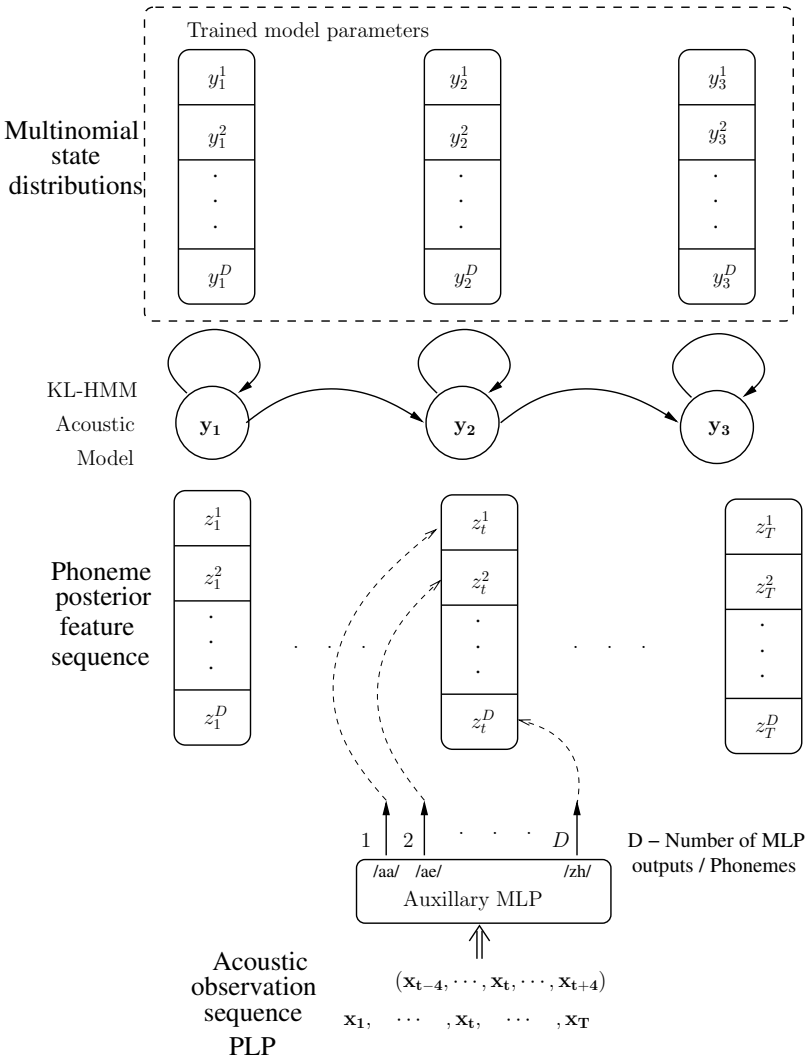

Figure A1. Illustration of KL-HMM system.

References

1. Sutton-Spence, R.; Woll, B. *The Linguistics of British Sign Language: An Introduction*; Cambridge University Press: Cambridge, UK, 1999. [CrossRef]
2. Cooper, H.; Holt, B.; Bowden, R. Sign Language Recognition. In *Visual Analysis of Humans*; Moeslund, T.B., Hilton, A., Kräger, V., Sigal, L., Eds.; Springer: London, UK, 2011; pp. 539–562. [CrossRef]

3. Adda-Decker, M.; Lamel, L. Multilingual Dictionaries. In *Multilingual Speech Proceesing*; Schultz, T., Kirchhoff, K., Eds.; Academic Press: Cambridge, MA, USA, 2006; Chapter 5, pp. 123–168.
4. Baus, C.; Gutiérrez, E.; Carreiras, M. The role of syllables in sign language production. *Front. Psychol.* **2014**, *5*, 1254. [CrossRef] [PubMed]
5. Boyes Braem, P.; Sutton-Spence, R. (Eds.) *The Hands Are the Head of the Mouth: The Mouth as Articulator in Sign Languages*; Signum: Hamburg, Germany, 2001.
6. Kaplan, R.; Kay, M. Regular models of phonological rule systems. *Comput. Linguist.* **1994**, *20*, 331–378.
7. Davel, M.; Barnard, E. Pronunciation prediction with Default&Refine. *Comput. Speech Lang.* **2008**, *22*, 374–393.
8. Dedina, M.; Nusbaum, H. PRONOUNCE: A program for pronunciation by analogy. *Comput. Speech Lang.* **1991**, *5*, 55–64. [CrossRef]
9. Pagel, V.; Lenzo, K.; Black, A. Letter to Sound Rules for Accented Lexicon Compression. In Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, 30 November–4 December 1998.
10. Bisani, M.; Ney, H. Joint-sequence Models for Grapheme-to-phoneme Conversion. *Speech Commun.* **2008**, *50*, 434–451. [CrossRef]
11. Wang, D.; King, S. Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields. *IEEE Signal Process. Lett.* **2011**, *18*, 122–125. [CrossRef]
12. Park, A.; Glass, J.R. Towards unsupervised pattern discovery in speech. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, San Juan, Puerto Rico, 27 November–1 December 2005; pp. 53–58.
13. Park, A.S.; Glass, J.R. Unsupervised Pattern Discovery in Speech. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 186–197. [CrossRef]
14. Varadarajan, B.; Khudanpur, S.; Dupoux, E. Unsupervised Learning of Acoustic Sub-word Units. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Columbus, OH, USA, 16–17 June 2008; pp. 165–168.
15. Jansen, A.; Church, K.; Hermansky, H. Towards spoken term discovery at scale with zero resources. In Proceedings of the Interspeech, Makuhari, Chiba, Japan, 26–30 September 2010; pp. 1676–1679.
16. Versteegh, M.; Thiollière, R.; Schatz, T.; Cao Kam, X.N.; Anguera, X.; Jansen, A.; Dupoux, E. The Zero Resource Speech Challenge 2015. In Proceedings of the Interspeech, Dresden, Germany, 6–10 September 2015.
17. Dunbar, E.; Cao, X.N.; Benjumea, J.; Karadayi, J.; Bernard, M.; Besacier, L.; Anguera, X.; Dupoux, E. The zero resource speech challenge 2017. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 323–330.
18. Lee, C.H.; Juang, B.H.; Soong, F.; Rabiner, L. Word recognition using whole word and subword models. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK, 23–26 May 1989; Volume 1, pp. 683–686.
19. Svendsen, T.; Paliwal, K.K.; Harborg, E.; Husoy, P.O. An improved sub-word based speech recognizer. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK, 23–26 May 1989; Volume 1, pp. 108–111.
20. Paliwal, K. Lexicon-building methods for an acoustic sub-word based speech recognizer. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 3–6 April 1990; pp. 729–732.
21. Svendsen, T.; Soong, F.; Purnhagen, H. Optimizing baseforms for HMM-based speech recognition. In Proceedings of the EUROSPEECH, Madrid, Spain, 18–21 September 1995.
22. Holter, T.; Svendsen, T. Combined optimisation of baseforms and model parameters in speech recognition based on acoustic subword units. In Proceedings of the ASRU, Santa Barbara, CA, USA, 17 December 1997; pp. 199–206.
23. Bacchiani, M.; Ostendorf, M. Joint lexicon, acoustic unit inventory and model design. *Speech Commun.* **1999**, *29*, 99–114. [CrossRef]
24. Singh, R.; Raj, B.; Stern, R.M. Automatic generation of subword units for speech recognition systems. *IEEE Trans. Speech Audio Process.* **2002**, *10*, 89–99. [CrossRef]
25. Hartmann, W.; Roy, A.; Lamel, L.; Gauvain, J. Acoustic unit discovery and pronunciation generation from a grapheme-based lexicon. In Proceedings of the ASRU, Olomouc, Czech Republic, 8–12 December 2013; pp. 380–385.

26. Lee, C.; Zhang, Y.; Glass, J.R. Joint Learning of Phonetic Units and Word Pronunciations for ASR. In Proceedings of the EMNLP, Seattle, WA, USA, 18–21 October 2013; pp. 182–192.
27. Razavi, M.; Rasipuram, R.; Magimai.-Doss, M. Towards Weakly Supervised Acoustic Subword Unit Discovery and Lexicon Development Using Hidden Markov Models. *Speech Commun.* **2018**, *96*, 168–183. [CrossRef]
28. Hanke, T. HamNoSys—Representing sign language data in language resources and language processing contexts. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, 26–28 May 2004; pp. 1–6.
29. Pitsikalis, V.; Theodorakis, S.; Vogler, C.; Maragos, P. Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In Proceedings of the IEEE CVPR Workshops, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1–6. [CrossRef]
30. Cooper, H.; Ong, E.; Pugeault, N.; Bowden, R. Sign language recognition using sub-units. *J. Mach. Learn. Res.* **2012**, *13*, 2205–2231.
31. Koller, O.; Ney, H.; Bowden, R. May the force be with you: Force-aligned signwriting for automatic subunit annotation of corpora. In Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (AFGR), Shanghai, China, 22–26 April 2013; pp. 1–6. [CrossRef]
32. Bauer, B.; Kraiss, K.F. Towards an Automatic Sign Language Recognition System Using Subunits. In *Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 64–75. [CrossRef]
33. Junwei, H.; George, A.; Alistair, S. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pattern Recognit. Lett.* **2009**, *30*, 623–633. [CrossRef]
34. Fang, G.; Gao, X.; Gao, W.; Chen, Y. A novel approach to automatically extracting basic units from Chinese sign language. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 23–26 August 2004; Volume 4, pp. 454–457. [CrossRef]
35. Theodorakis, S.; Pitsikalis, V.; Maragos, P. Model-level data-driven sub-units for signs in videos of continuous Sign Language. In Proceedings of the IEEE ICASSP, Dallas, TX, USA, 14–19 March 2010; pp. 2262–2265. [CrossRef]
36. Sako, S.; Kitamura, T. Subunit modeling for Japanese sign language recognition based on phonetically depend multi-stream hidden Markov models. In *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 548–555.
37. Miller, G.A. *The Science of Words*; W. H. Freeman and Company: New York, NY, USA, 1996.
38. Bhattacharyya, A. On a Measure of Divergence between Two Multinomial Populations. *Sankhyā Indian J. Stat.* **1946**, *7*, 401–406.
39. Aradilla, G.; Vepa, J.; Bourlard, H. An acoustic model based on Kullback-Leibler divergence for posterior features. In Proceedings of the ICASSP, Honolulu, HI, USA, 15–20 April 2007.
40. Aradilla, G.; Bourlard, H.; Magimai.-Doss, M. Using KL-based acoustic models in a large vocabulary recognition task. In Proceedings of the Interspeech, Brisbane, Australia, 22–26 September 2008.
41. Magimai.-Doss, M.; Rasipuram, R.; Aradilla, G.; Bourlard, H. Grapheme-based Automatic Speech Recognition using KL-HMM. In Proceedings of the Interspeech, Florence, Italy, 27–31 August 2011.
42. Razavi, M.; Rasipuram, R.; Magimai.-Doss, M. Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework. *Speech Commun.* **2016**, *80*, 1–21. [CrossRef]
43. Rasipuram, R.; Magimai.-Doss, M. Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model. *Speech Commun.* **2015**, *68*, 23–40. [CrossRef]
44. Tornay, S.; Razavi, M.; Camgoz, N.C.; Bowden, R.; Magimai.-Doss, M. HMM-based Approaches to Model Multichannel Information in Sign Language inspired from Articulatory Features-based Speech Processing. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019.
45. Pignat, E.; Calinon, S. Learning adaptive dressing assistance from human demonstration. *Robot. Auton. Syst.* **2017**, *93*, 61–75. [CrossRef]
46. Bohner, M.; Wintz, N. The Linear Quadratic Tracker on time scales. *Int. J. Dyn. Syst. Differ. Equ.* **2011**, *3*. [CrossRef]

47. Hermansky, H. Perceptual Linear Predictive (PLP) Analysis of Speech. *J. Acoust. Soc. Am.* **1990**, *57*, 1738–1752. [CrossRef]
48. Pitrelli, J.F.; Fong, C.; Wong, S.H.; Spitz, J.R.; Leung, H.C. PhoneBook: A phonetically-rich isolated-word telephone-speech database. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Detroit, MI, USA, 9–12 May 1995; Volume 1, pp. 101–104. [CrossRef]
49. Dupont, S.; Bourlard, H.; Deroo, O.; Fontaine, V.; Boite, J.M. Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on 'Phonebook' and Related Improvements. In Proceedings of the ICASSP, Munich, Germany, 21–24 April 1997.
50. Rabiner, L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **1989**, *77*, 257–286. [CrossRef]
51. Bourlard, H.; Morgan, N. *Connectionist Speech Recognition: A Hybrid Approach*; Kluwer Academic Publishers: Norwell, MA, USA, 1993.
52. Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Liu, X.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; et al. *The HTK Book*; Cambridge University Engineering Department: Cambridge, UK, 2002.
53. Johnson, D.; Ellis, D.; Oei, C.; Wooters, C.; Faerber, P.; Morgan, N.; Asanovic, K. ICSI Quicknet Software Package. 2004. Available online: http://www.icsi.berkeley.edu/Speech/qn.html (accessed on 8 January 2018).
54. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
55. Ong, S.C.W.; Ranganath, S. Automatic Sign Language Analysis: A Survey and the Future Beyond Lexical Meaning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 873–891. [CrossRef] [PubMed]
56. Ebling, S.; Camgoz, N.C.; Braem, P.B.; Tissi, K.; Sidler-Miserez, S.; Stoll, S.; Hadfield, S.; Haug, T.; Bowden, R.; Tornay, S.; et al. SMILE Swiss German sign language dataset. In Proceedings of the Language Resources and Evaluation Conference, Miyazaki, Japan, 7–12 May 2018.
57. Camgöz, N.C.; Kındıroğlu, A.A.; Akarun, L. Sign Language Recognition for Assisting the Deaf in Hospitals. In Proceedings of the Human Behavior Understanding: 7th International Workshop, Amsterdam, The Netherlands, 16 October 2016; Springer International Publishing: Cham, Switzerland, 2016; pp. 89–101. [CrossRef]
58. Koller, O.; Ney, H.; Bowden, R. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
59. Jachova, Z.; Olivera, K.; Karovska Ristovska, A. Differences between American Sign Language (ASL) and British Sign Language (BSL). *J. Spec. Educ. Rehabil.* **2008**, *9*, 41–52. [CrossRef]
60. King, S.; Frankel, J.; Livescu, K.; McDermott, E.; Richmond, K.; Wester, M. Speech production knowledge in automatic speech recognition. *J. Acoust. Soc. Am.* **2007**, *121*, 723–742. [CrossRef] [PubMed]
61. Rasipuram, R.; Magimai.-Doss, M. Articulatory feature based continuous speech recognition using probabilistic lexical modeling. *Comput. Speech Lang.* **2016**, *36*, 233–259. [CrossRef]
62. Wrench, A.; Richmond, K. Continuous Speech Recognition Using Articulatory Data. In Proceedings of the ICSLP, Beijing, China, 16–20 October 2000.
63. Richmond, K.; Hoole, P.; King, S. Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus. In Proceedings of the Interspeech, Florence, Italy, 27–31 August 2011.