RECOGNITION WITH RAW CANONICAL PHONETIC MOVEMENT AND HANDSHAPE SUBUNITS ON VIDEOS OF CONTINUOUS SIGN LANGUAGE

Stavros Theodorakis, Vassilis Pitsikalis, Isidoros Rodomagoulakis and Petros Maragos

School of E.C.E., National Technical University of Athens, Greece

ABSTRACT

The visual processing of Sign Language (SL) videos offers multiple interdisciplinary challenges for image processing and recognition. Based on tracking and visual feature extraction, we investigate SL visual phonetic modeling by exploiting statistical subunit (SU) models of movement-position and handshape. We further propose a new framework to construct a data-driven lexicon that retains phonetics' movement information and to perform automatic recognition of continuous SL videos. We construct phonetically meaningful transition SU, named as *raw canonical phonetic subunits* (SU-CanRaw). Then, we integrate via a Hidden Markov Model multistream scheme the SU-CanRaw extended for both hands, with handshape SU, based on our previous work on Affine-invariant Shape-Appearance Models. By applying the all-inclusive framework on continuous SL videos, we automatically generate a data-driven lexicon that can be further exploited, for automatic analysis of SL corpora, and continuous SL recognition. The recognition experiments, conducted on a newly acquired continuous SL corpus, lead to promising results.

Index Terms— Automatic Sign Language (SL) Recognition, Phonetic Models, Movement Handshape Integration, Automatic Data-Driven Phonetic Lexicon, Greek SL.

1. INTRODUCTION

Sign Languages (SL's) are languages employing manual and non-manual visual patterns, and their automatic processing poses several interdisciplinary challenges [12], e.g. the tracking, feature extraction and statistical visual-phonetic modeling. The lack of phonetic transcriptions, standardized phonetic models and lexica for SL corpora render continuous Sign Language Recognition (SLR) quite difficult [19, 16]. SLR tasks are found even more demanding due to the variability of continuous signing characteristics and the multiple information streams, as for instance handshape and movement. Next, we present a framework for continuous SLR incorporating a new visual-phonetic modeling approach and constructing a data-driven lexicon without any prior phonetic information.

Speech recognition systems require a phoneme set, a phonetic lexicon, and an annotated data corpus. SLR is required to deal with many *new* issues compared to speech [12, 17], however the above ingredients are somehow indispensable. Despite the interdisciplinary SL research progress such things, as a phonetic lexicon, are not yet standard. Nor is it easy to produce precise phonetic SL corpora transcriptions given the multiple parallel cues. Phonetic transcriptions refer to the annonation of visual events. A phoneme in sign corresponds to the basic components of the multiple cues, e.g. a basic movement, or handshape: see for instance the downwards movement or the V-like handshape in Fig. 1(a) for sign SATISFACTION. The lack of well-defined and accepted computational phonemes has been dealt by either employing sign-level models [2, 3], or data-driven methods [4, 5, 6, 7, 8]. However the latter result in linguistically meaningless subunits (SU). A few approaches have incorporated linguistic knowledge such as [9, 3, 10]. Recently in [11] sequential phonetic descriptions are mapped with statistical phonetic models advancing towards a direction that is by-default available for speech.

Another issue for SL is the articulation of multiple information streams. Their integration is still open for automatic SLR [12]. From the linguistic viewpoint there is an ongoing evolution of concepts on the relations of the multiple streams [13, 14]. Integration schemes such as parallel Hidden Markov Models (HMM) given manual transcriptions have been presented, in [15] . Multiple cues are combined in [10] for isolated sign recognition. Another aspect concerns continuous SLR [16, 18, 19, 20] and issues such as coarticulation and movement epenthesis. Nested dynamic programming is employed in [18] to handle movement epenthesis. Transition-movement models are employed for large-vocabulary continuous SLR [19]; [20] presents a threshold model based on conditional random fields.

In this article, we present a SU-based statistical framework, for the automatic recognition of continuous SL videos, that consists of: 1) The visual processing and feature extraction [21]. 2) The statistical SU construction. 3) The automatic unsupervised lexicon construction and 4) continuous recognition. We introduce a new method for statistical visual SU models referred to as *raw canonical phonetic subunits* (SU-CanRaw). These are built 1) by uniformly sampling the feature space and constructing statistical HMM models that carry *by-construction* phonetic information, and 2) by encapsulating data-driven phonetic information of dynamic and static parts (as in [8]) to handle sequentiality of movements and postures respectively. In addition, we enhance the phonetic modeling via a simple tying, sharing of model parameters, scheme for the systematic incorporation of the non-dominant hand. We also integrate handshape information by adapting our previous works on feature extraction [22], on SU construction [21] and on preliminary movement-handshape integration [23]. Finally, we generate a data-driven sign-level lexicon which retains phonetics' movement information due to the SU-CanRaw models, and despite the lack of phonetic corpus transcriptions. This is in contrast to data-driven SUs. We should also stress that the phonetics-based SU presented in [11] could not be employed as is, due to the lack of precise phonetic data annotations. The whole framework is applied on a newly acquired continuous SL corpus leading to promising results.

2. DATA AND VISUAL PROCESSING

The *Greek Sign Language (GSL) Corpus* contains data of multiple tasks [1] [1]. Figure 1 shows a data sample. For the segmenta-

This research work was supported by the EU under the research program Dictasign with grant FP7-ICT-3-231135.

[1]We focus on data from Task-4. The processed videos have a resolution of 720x576 pixels at 25 fps and sign-level transcriptions.

(a) SATISFACTION (b) T (c) LOOK (d) T (e) AIRPLANE (f) T (g) SHOCK

Fig. 1. Sample of continuous signing utterance: "SATISFACTION LOOK AIRPLANE SHOCK"; epenthesis transitions (T) are shown in between signs due to coarticulation. For each transition the first and the last frames of each are superimposed with indicative arrows.

tion and detection of the signer's hands and head in the GSL corpus we employ a skin color model with a Gaussian Markov Model (GMM) and morphological processing to enhance skin detection and face/hand region segmentation (as in [21]). For tracking we employ linear prediction, and template matching to disambiguate occlusions. The movement feature vector consists of the 2D coordinates of the hand, the instantaneous direction, and the velocity. For the handshape feature extraction we employ an affine-invariant modeling of hand shape-appearance images that model the handshapes without any landmark points, employing a linear combination of variation images followed by affine transformations. The fitting is based on optimization, resulting on the estimated variation coefficients, i.e. the handshape features as in [22].

3. MOVEMENT RAW CANONICAL SUB-UNIT MODELS

3.1. Why Go Raw ?

SU-CanRaw Statistical Sub-Unit Models. The Phonetics-Based SU of [11] can be employed when precise phonetic transcriptions exist. In their absence and since the low-level phonetic annotation is time consuming, the alternatives are either to work towards phonetic adaptation for continuous signing, or to construct new models that account for the missing phonetic information. Next, we investigate the latter. We introduce the Raw Canonical SU (SU-CanRaw) statistical models that by default carry phonetics information.

Feature Space Issues. A usual issue when training models is the unequally or sparsely populated feature space. This can hurt either data-driven or phonetic-based approaches. Especially for the phonetics-based case [11] due to the large number of phonetic labels (HamNoSys[2]symbols), it is common that phonetic models are barely populated or not at all. Thus during model training there shall be both missing or poorly trained models. This fact affects recognition as well. With SU-CanRaw models this situation is dealt by deterministically populating the feature space with models. Herein we deal with the simplest case in which the feature space is uniform; i.e. the case of transition straight and curved movements.

3.2. How to costruct them

The phonetic movement related transcriptions (corresponding to HamNoSys symbols) are characterized by symmetry. For instance the straight lines sample the 3D directions in the signing space [3]. We take advantage of the above and deterministically *define* statistical models that correspond to these equally spaced initializations.

[2]The Hamburg SL Notation System [24]: a "phonetic" transcription system employed for SL phonetic description.

[3]We consider the straight and curved transitions. In spite of the existence of more HamNoSys symbols, the accounted ones provide enough variability to describe a variety of movements and signs. Signing space refers to the physical 3D space in which the hands move.

Transition Raw Canonical SU Models. For the transition models we employ HMM models [26] considering the curved and straight transitions. We uniformly partition the hand's transition direction feature space generating all the different model initializations for the straight and curved types. The straight lines partitioning is illustrated for the spatial signing space in Fig. 2 (a), normalizing the transitions at the same initial point $(0, 0)$. We then construct a 5-state HMM for each transition on the direction feature space. The mean parameters for each HMM state correspond to the point markers shown in Fig. 2 (a) after the equal directions partitioning. Equal model variance is employed for each state requiring non-overlapping gaussians. This is illustrated in the models' states Fig. 2(b) of each transition. An example of these transition SU-CanRaw is shown in Fig. 3(b): there the transition in the sign SHOCK corresponds to the T2 transition SU-CanRaw model Fig. 2(b) green dotted line (i.e. an upper-right transition). All the above concern the direction feature and require no training. We incorporate statistics on velocity in Section 3.3. Similarly we construct curved SUs transitions.

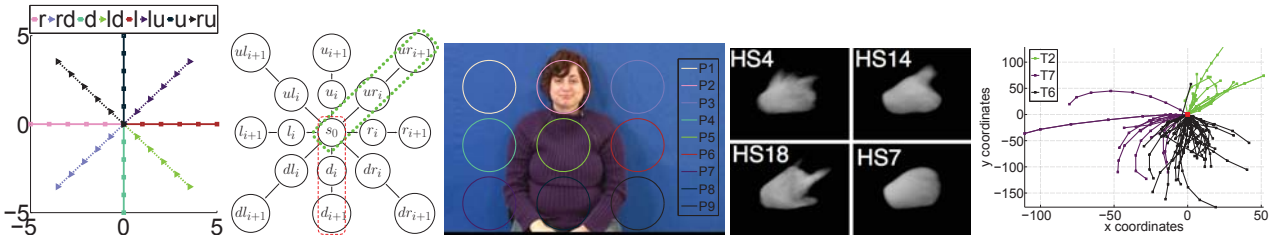
Posture SU Models. For the postures we uniformly partition the 2D sign space creating the canonical posture models. Figure 2(c) shows the partitioning in the 2D sign space. Then we construct a GMM for each posture $(P_i, i = 1, .., 9)$ employing as feature the x,y coordinates. These models are not "canonical" as the transition models w.r.t. HamNoSys. They are rather initialized in a uniform way. Posture SU model examples are illustrated in Fig. 3(d,f); these correspond to the $(P4, P5)$ posture SU's of sign SATISFACTION.

3.3. Adding Dynamic-Static Data-Driven Statistics

The main characteristics prevailing during SL articulation are multiple streams and sequentiality. Driven by the Movement-Hold sequential structure [14] and given the lack of phonetic information, it is essential to enhance the SU-CanRaw models so that they account for sequentiality. For this, we employ the Dynamic-Static data-driven method [8] on unsupervised detection of sequential Postures (Static) and Transitions (Dynamic). For the separation of the dynamic/static parts, we employ the velocity feature. Dynamic (D) parts correspond to movements and Static (S) to non-movements. Considering that the transition SU-CanRaw models are built on the direction feature space, they contain no information on velocity. We incorporate such dynamics by increasing with an extra stream the SU-CanRaw models. In this way the final SU models contain the statistics on direction that partition the spatial domain and at the same time they encapsulate data-driven information that has been efficient for the sequential detection of dynamic (transitions) vs. static (postures). Similarly we enhance the posture models.

4. INCORPORATION OF NON-DOMINANT HAND

The multiple information sources are integrated via a multistream HMM scheme. From the SL articulation perspective a single hand is



(a) (b) (c) (d) (e)

Fig. 2. Canonical SU-CanRaw movement SU (top row): a) Straight transitions' partitioning (right (r), left (l), up (u), down (d)) in the signing space. b) Corresponding SU-CanRaw HMM Models. c) Uniform distributed posture SU. d) Samples of handshape SU.e) Data-Driven SU.

considered as dominant (right hand) constructing the major manual phonetic part of signs. However, the non-dominant hand (left hand) can contribute either as a supplementary dominant component or as another place-of-articulation. However from the movement phonetics viewpoint we do not restrict a priori the possible movements of the non-dominant hand. A crucial point is that the movement SUs (of any hand) could be tied with a standard set of basic movement models (the SU-CanRaw models) independent to which hand is performing this movement, and whether both or a single hand takes part. By tying we refer to the sharing of the statistical parameters of the underlying distributions, while on each update all models are updated. Thus, we replicate the main set of movement SU-CanRaw phonetic models via tying for the movements of either hand. Building on the same concept we also construct transitions-of-both-hands SU named hereafter as MB-SU models by taking the product of possible combinations. An example of these MB-SU models is illustrated in Fig. 3(e) which corresponds to a transition of both hands for the sign SATISFACTION. Both hands perform the same downward transition corresponding to the T7 transition SU-CanRaw model (see Fig.2(b), red dotted line). Thus a combination of these transitions constructs a MB-SU model (MB-T7-T7).

5. MOVEMENT-POSITION HANDSHAPE INTEGRATION

Herein we exploit the *dynamic-static* nature of the SU-CanRaw models by considering Handshapes (HS) only during postures. We partition via K-means the feature space of HS features (Sec. 2) and train a GMM model for each cluster that corresponds to a data-driven HS SU (HSi). Indicative samples of the mean shape-appearance reconstructed images of the centroids for the HS SU are shown in Fig. 2(d). These are intuitive as each HS SU corresponds to a different hand configuration. In addition the HS employed in each sign and its corresponding HS SU are similar: for instance HS4, HS14, HS18 correspond to signs SHOCK, SATISFACTION and AIRPLANE respectively (see Figs. 1, 3, 2(d)). This scheme (as in [23]) results on *fused SU models* of Posture and HS-SU Models (P_i-HS_j), i and j correspond to the single-cue SU index of postures (P) and handshapes (HS). Fused SU model examples are illustrated in Fig. 3(a,c) and Fig. 3(d,f) for signs SHOCK and SATISFACTION respectively. For instance, P4-HS4 SU is a combination of the P4 posture SU model (Fig.2(c)) and of the HS4 SU model (Fig.2(d)).

6. LEXICON AND CONTINUOUS RECOGNITION

Data-Driven Lexicon Construction. Employing resources presented in the previous sections we segment each sign instace via a model-based segmentation. For this we do not incorporate any phonetics information in terms of intra-sign ground-truth phonetic annotation (in contrast to [11]). Thus in these terms it is data-driven. Nevertheless, via the inherent phonetics incorporated by the SU-CanRaw movement models the resulting lexicon retains this phonetic information. Specifically, we concatenate the SU-CanRaw models in a network and decode via HMMs each feature sequence via the Viterbi algorithm, generating a sequence of SU labels with their start/end frames. Applying the above for all signs (since we consider their boundaries) leads to the formation of a lexicon with intra-sign segmentation boundaries per sign. The units of this lexicon may contain either transition based SUs (SU-CanRaw) or posture+handshape SUs. Figures 3(a-c) and Fig. 3(d-f) show the decomposition of two signs, SHOCK and SATISFACTION respectively into the SU they consist. Sign SATISFACTION (Figs. 1(c), 3(a-c)) consists of 3 SUs: A posture-handshape SU (P4-HS14), a transition-of-both-hands SU (MB-T7-T7) and a posture-handshape SUs (P5-HS14).

Utterance-level Continuous Recognition. We then consider the continuous utterance-level stream *without seeing the corresponding sign boundaries* (see Fig. 1). Our resources are: 1) the constructed SU models which constitute the phoneme-set together with their statistical SU models; 2) the modeling architecture; 3) the unsupervised data-driven lexicon. Thus, we have accounted for the missing basic ingredients of the considered recognition task.

7. EXPERIMENTS

Experimental Configuration: The experiments are conducted on data from Task-4, Signer-12B of the GSL corpus [1], which contains 52 utterances, 142 different signs and 461 total sign instances. Each utterance consists of 10 signs on average. For **evaluation** we employ the metrics: *Sign Correct* $= (N-D-S)/N{\cdot}100\%$ and *Sign Accuracy* $= (N-D-S-I)/N\cdot 100\%$; N corresponds to the total number of signs and D, S, I to Deletion, Substitution and Insertion errors.

Data-Driven (SU-DD) vs. SU-CanRaw SU: SU-DD SU correspond to dynamic and static subunits that have been constructed automatically employing 1) a model-based segmentation into movements and non-movements based on velocity and 2) an unsupervised clustering of segments (see [8, 25] for more details). The SU-DD approach has been shown to have advantages within the set of data-driven based SU approaches ([4, 5]). It thus consists a decent performing baseline. For the SU-DD construction we use the same number of dynamic and static clusters as in the SU-CanRaw (16 and 9 clusters for the dynamic and static SUs respectively). For the training of the SU-DD we employ a training set in contrast to the SU-CanRaw which have been created constructively without the need of a training set. As observed in Figs. 4(a,b) the recognition results are similar. Therefore, with SU-CanRaw we can still obtain similar performance and maintain the advantage of SU-CanRaw models i.e. the mapping of the SUs to the phonetic labels. For instance, we present the transition SUs that correspond to the SU-DD T2, T7, T6 in Fig. 2(e). By comparing these with the corresponding transitions of the SU-CanRaw model we observe that they are more complex with increased variance and without a clear separation between

(a) P4-HS4 (b) T2 (c) P5-HS4 (d) P4-HS14 (e) MB-T7-T7 (f) P5-HS14

Fig. 3. Example of sign SHOCK decomposed into sub-units P4-HS4 T2 P5-HS4 (a-c). Example of sign SATISFACTION decomposed into sub-units P4-HS14 MB-T7-T7 P5-HS14 (d-f). P_i correspond to posture SU models, HS_i to handshape SU models, T_i to transition SU-CanRaw models and MB-T_i-T_j transition-of-both-hands SU models.

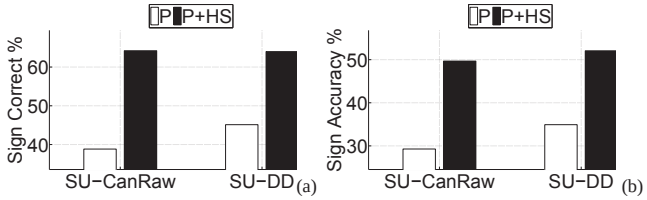


Fig. 4. Continuous SL Recognition: Comparison of the proposed framework with 1) SU-CanRaw vs. Data-Driven SUs (SU-DD) 2) With handshape information (P+HS) and without (P).

neighboring models. The variance may results in an advantage for recognition since they are adapted on the *exact* data. However there doesn't exist a correspondence to the clear directions of phonetic annotations (Fig. 2(e)). Finally, note that the SU-CanRaw framework concerning transitions-postures *sees actually no data* appart from a 10% of the dataset to train the dynamic-static statistics.

Handshape Integration: After incorporating the handshape stream the recognition performance increases at least by 17% and 18% in % Correct and Accuracy respectively (Fig. 4). An example of the sign decoding of 2 utterances (the first corresponds to utterance in Fig. 1) with (P+HS) and without (P) employing handshape information; REF corresponds to the ground truth sign transcriptions. As observed, by incorporating the handshape information stream more signs are recognized in both cases.

REF	SATISFACTION	LOOK	AIRPLANE	SHOCK
P	OBLIGATION	X	X	SHOCK
P+HS	SATISFACTION	X	AIRPLANE	SHOCK
REF	ARRIVE	SATISFACTION	JOURNEY	SEE
P	ARRIVE	FOLLOW	JOURNEY	X
P+HS	ARRIVE	SATISFACTION	JOURNEY	SEE

8. CONCLUSIONS

The presented new framework for automatic recognition of continuous SL investigates various aspects of visual-phonetic modeling, by building a higher level of statistical modeling over new structural units (SU-CanRaw) of visual movement-transitions on manual articulation. This retains movement phonetics' information despite the lack of any phonetic lexicon/corpus annotation. This framework together with handshape integration is applied on a demanding continuous SL task with promising results; further experiments are to consider multiple signers, increased vocabulary and non-manual cues.

9. REFERENCES

[1] `http://www.dictasign.eu`.

[2] P. Buehler, M. Everingham, and A. Zisserman, "Learning sign language by watching TV (using weakly aligned subtitles)," in *CVPR*,2009, pp. 2961–2968.

[3] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, "A linguistic feature vector for the visual interpretation of sign language," in *ECCV*, 2004.

[4] B. Bauer and K. F. Kraiss, "Towards an automatic sign language recognition system using subunits," in *Proc. of Int'l Gesture Workshop*, 2001, pp. 64–75.

[5] G. Fang, X. Gao, W. Gao, and Y. Chen, "A novel approach to automatically extracting basic units from chinese sign language," in *ICPR*, 2004.

[6] J. Han, G. Awad, and A. Sutherland, "Modelling and segmenting subunits for sign language recognition based on hand motion analysis," *Pat. Rec. Lett.*, vol. 30, no. 6, pp. 623–633, 2009.

[7] P. Yin, T. Starner, H. Hamilton, I. Essa, and J. Rehg, "Learning the basic units in american sign language using discriminative segmental feature selection," in *Proc. ICASSP*, 2009, pp. 4757–4760.

[8] V. Pitsikalis, S. Theodorakis, and P. Maragos, "Data-driven sub-units and modeling structure for continuous sign language recognition with multiple cues," in *LREC Wrk Representation and Proc. of SL: Corpora and SL Technologies*, 2010.

[9] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of american sign language," *CVIU*, vol. 81, no. 3, pp. 358–384, 2001.

[10] L. Ding and A. Martinez, "Modelling and recognition of the linguistic components in american sign language," *Im. and Vis. Comp.*, 27, pp. 1826–1844, 2009.

[11] V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos, "Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition," in *CVPR Workshop Gesture Recognition*, 2011.

[12] U. Agris, J. Zieren, U. Canzler, B. Bauer, and K. F. Kraiss, "Recent developments in visual sign language recognition," *Univ. Acc. Inf. Soc.*, 6, pp. 323–362, 2008.

[13] WC Stokoe, "Sign language structure," *Annual Review of Anthropology*, vol. 9, no. 1, pp. 365–390, 1980.

[14] S. K. Liddell and R. E. Johnson, "American sign language: The phonological base," *Sign Language Studies*, vol. 64, pp. 195 – 277, 1989.

[15] C. Vogler and D. Metaxas, "Handshapes and movements: Multiple-channel american sign language recognition," in *Gesture Workshop*, 2003, pp. 247–258.

[16] P. Dreuw, J. Forster, T. Deselaers, and H. Ney, "Efficient approximations to model-based joint tracking and recognition of continuous sign language," in *Proc. IEEE Int'l Conf. on Autom. Face & Gesture Rec.*, 2008, pp. 1–6.

[17] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech recognition techniques for a sign language recognition system," in *Interspeech*, 2007.

[18] R. Yang, S. Sarkar, and B. Loeding, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," *IEEE Trans. PAMI*, vol. 32, no. 3, pp. 462–477, 2010.

[19] G. Fang, W. Gao, and D. Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," *IEEE Trans. SMC*, 37, 2007.

[20] H. D. Yang, S. Sclaroff, and S. W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *IEEE Trans. PAMI*, 31, 2009.

[21] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos, "Hand tracking and Affine Shape-Appearance Handshape Sub-Units in Continuous Sign Language recognition," in *Workshop on Sign, Gesture and Activity (SGA), ECCV*, 2010.

[22] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos, "Affine-invariant modeling of shape-appearance images applied on sign language handshape classification," in *Proc. Int'l Conf. on Image Processing*, 2010.

[23] S. Theodorakis, V. Pitsikalis, and P. Maragos, "Advances in dynamic-static integration of movement and handshape cues for sign language recognition," in *Gesture Workshop (GW-2011), Greece, Athens, May 2011*, 2011.

[24] S. Prillwitz, R. Leven, H. Zienert, R. Zienert, T. Hanke, and J. Henning, "HamNoSys. Version 2.0," *Int'l Studies on SL and Communication of the Deaf*, 1989.

[25] S. Theodorakis, V. Pitsikalis, and P. Maragos, "Dynamic vs. Static Sub-Unit Modeling for Continuous Sign Language Recognition with Movement-Position Cues," *IEEE Trans. Sys. Man Cyb. Part B*, under review, 2012.

[26] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, pp. 257–286, 1989.