Vetenskapsrådet

Application

2015-03-31

2015-05202	Salvi, Giampiero	NT-2

Information about applicant

Name: Giampiero Salvi

Doctorial degree: 2006-11-02

Birthdate: 19730501

Academic title: Docent

Gender: Male

Employer: Kungliga Tekniska högskolan

Administrating organisation: Kungliga Tekniska högskolan

Project site: TMH, Tal, musik och hörsel

Information about application

Call name: Forskningsbidrag Stora utlysningen 2015 (Naturvetenskap och teknikvetenskap)

Type of grant: Projektbidrag

Focus: Fri

Subject area:

Project title (english): Recurrent Pattern Discovery in Unsegmented Sequences

Project start: 2016-01-01

Project end: 2019-12-31

Review panel applied for: NT-2, NT-14, NT-1

Classification code: 10208. Språkteknologi (språkvetenskaplig databehandling), 10207. Datorseende och robotik (autonoma system), 10201. Datavetenskap (datalogi)

Keywords: unsupervised pattern discovery, automatic speech understanding, human activity recognition, language acquisition, segmentation

Funds applied for

Year:	2016	2017	2018	2019
Amount:	1,407,000	1,445,000	1,552,000	1,671,000

Descriptive data

Project info

Project title (Swedish)*

Obevakad segmentering och klassificering av återkommande mönster i sekvensiell data

Project title (English)*

Recurrent Pattern Discovery in Unsegmented Sequences

Abstract (English)*

This project aims at developing general purpose methods for the discovery of recurrent patterns in unsegmented sequences. We will also apply these methods to the discovery of fundamental units in speech and in human actions and gestures analysis.
Machines still lag behind humans when it comes to learning from unsegmented sequential inputs. During early development, humans learn to segment running speech into phonemes, words and phrases, and to split human actions or communicative gestures into fundamental constituent movements. It is natural for us to perceive a piece a music as composed by musical phrases (sequences of notes with a complete musical sense). In machine learning, however, the problem of automatically segmenting and classifying streams of sensory inputs without annotations still does not have a definite solution. We will propose a solution to this problem based on infinite latent variable models. We will show how the methods developed in the project can extract, without annotations, fundamental units of speech that are feasible modelling blocks for speech recognition. We will also test these methods on a number of scenarios in computer vision, including human activity, gesture, and sign language recognition.
We expect the results of this project to be relevant to many other fields, such as robotics, econometrics, bioinformatics, to name some. In order to encourage this, we will make all the developed methods available to the research community as free software.

Popular scientific description (Swedish)*

Vi vill låta maskiner lära sig på ett liknande sätt som människor gör. För oss människor är det naturligt att upptäcka mönster i ljuden som vi hör eller rörelserna som vi ser. Vi hittar betydelsebärande enheter i det kontinuerliga flödet av information som vi utsätts för. Det är, till exempel, naturligt för oss människor att uppfatta ett stycke musik som komponerat av musikaliska fraser (sekvenser av toner med en komplett musikalisk bemärkelse). Däremot, är det svårt för nuvarande maskiner att lära sig från sekvenser av observationer om vi inte hjälper dem med transkriptioner av sekvensernas innehåll. Vi kommer i projektet att utveckla och tillämpa maskininlärningsmetoder som kan hitta betydelsefulla segment i ljud och rörliga bilder. Vi kommer att testa våra metoder med ljudinspelningar av tal och videoinspelningar av mänskliga rörelser, till exempel, människor som lagar mat, eller som försöker kommunicera med gester, eller med teckenspråk.
Samma metoder kommer förhoppningsvis att vara relevanta även i helt olika område, till exempel robotik, ekonometri, beräkningsbiologi, med mera. De kommer att kunna hjälpa analyster hitta speciella händelser i ekonomin, eller biologer hitta sekvenser i DNA. I robotik, kommer metoderna att bidra till det långsiktiga målet att utveckla verkligt autonoma system, som kan fortsätta lära sig under hela sina liv, och anpassa sig till situationer som utvecklaren inte hade förutsett. Vi kommer, därför, att dela ut våra metoder så att de kan ge bästa möjliga nytta för samhället.

Project period

Number of project years*

4

Calculated project time*

2016-01-01 - 2019-12-31

Classifications

Select a minimum of one and a maximum of three SCB-codes in order of priority.

Select the SCB-code in three levels and then click the lower plus-button to save your selection.

SCB-codes*

1. Naturvetenskap > 102. Data- och informationsvetenskap (Datateknik) > 10208. Språkteknologi (språkvetenskaplig databehandling)

1. Naturvetenskap > 102. Data- och informationsvetenskap (Datateknik) > 10207. Datorseende och robotik (autonoma system)

1. Naturvetenskap > 102. Data- och informationsvetenskap (Datateknik) > 10201. Datavetenskap (datalogi)

Enter a minimum of three, and up to five, short keywords that describe your project.

Keyword 1*

unsupervised pattern discovery

Keyword 2*

automatic speech understanding

Keyword 3*

human activity recognition

Keyword 4

language acquisition

Keyword 5

segmentation

Research plan

Ethical considerations

Specify any ethical issues that the project (or equivalent) raises, and describe how they will be addressed in your research. Also indicate the specific considerations that might be relevant to your application.

Reporting of ethical considerations*

Ethical aspects are not relevant for this project. We will use existing recordings of human behavior (speech and human actions) that have already addressed the problem of protecting the subjects' privacy.

The project includes handling of personal data

No

The project includes animal experiments

No

Account of experiments on humans

No

Research plan

Recurrent Pattern Discovery in Unsegmented Sequences

1 Purpose and aims

In our previous research, we gathered valuable experience on modelling and imitating aspects of human language acquisition with computational Machine Learning methods. It has become evident to us that one of the main areas in which humans clearly outperform machines is the ability to learn from unconstrained sequences of inputs, i.e., sequences of patterns where neither the segmentation, nor the defining properties of the patterns are known in advance. For example, contrary to machines, humans learn to segment speech into words and phrases without any annotations. This is not limited to language acquisition: we effortlessly segment human actions into meaningful units, or perceive music as being composed by phrases (groupings of notes with a complete musical sense), to make other examples. The ability to group sequential inputs into meaningful segments is at the basis of our capacity to adapt to new situations and learn from an apparently unstructured environment in continuous motion.

In order to provide autonomous systems with this important ability, we propose the use of advanced Machine Learning methods that do not only optimise the parameters of a predefined model, as in traditional Machine Learning. On the contrary, these methods will be able to cope with unexpected inputs, and to discover patterns that were not, to large extent, anticipated by the model designer. This will lead to the automatic discovery of constituent building blocks for the problem at hand, may they be linguistic units for Automatic Speech Recognition or fundamental human movements for Human Activity and Gesture Recognition.

The main objective of this project is applying advance machine learning methods to the unsupervised discovery of recurrent patterns in unsegmented sequences. We will show that it is possible to automatically acquire linguistic constituents such as words and sub-word units from running speech. We will also show the generality of those methods by applying them to the area of human activity/gesture recognition. The ambition is to develop a general framework that can be applicable to a vast range of practical problems, in areas such as robotics, computer vision, econometrics, and bioinformatics.

1.1 Problem definition and Motivation

Figure 1 shows the waveform and spectrogram (distribution of energy in time and frequency) for a segment of child-directed speech where a mother talks to her child about some toys they are playing with. The new-born child's task is to learn to segment the signal into chunks that bear some meaning. The speech signal is produced by the mother without any interruption (the white vertical bars that are visible in the spectrogram are due to the physics of speech production and do not usually correspond to word segmentation). In order to illustrate and explain the fundamental nature of the problem faced by the new-born child, we will in the following consider sequences of discrete symbols such as those in Table 1. The arguments,

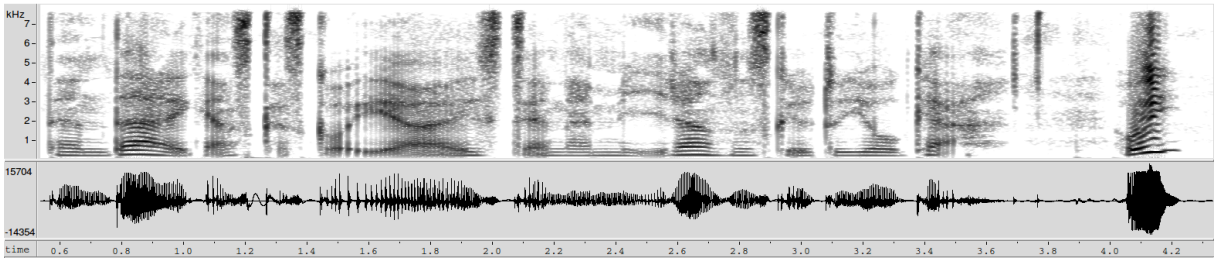

Figure 1: An example of child-directed speech

class 1	class 2	class 3
abababab	aaaaaaabbbbbbb	bbbbbbbbaaaaaaaaaa
ababab	aaaabbbb	bbbbaaaaa
abababababababab	aabb	bbbbbbbbbbbbbaaaaaaaaaa

Table 1: Classification of variable-length sequences

however, are as valid for sequences of continuously varying quantities such as the example in Figure 1.

In many machine learning problems, the observations we want to analyse can be expressed with a fixed-length representation. This is the case for all those problems where we can measure a number of attributes for each example, and those attributes are the same for all examples we consider. When we consider sequences of observations, however, resorting to a fixed-length representation often causes loss of information. As an illustration, consider the simple examples in Table 1 containing sequences over the alphabet {a,b}. If we want to build a classifier for this kind of observations, we need to apply specific methods (see Section 2) because the observations have different lengths. This is true also in cases where we want to perform unsupervised clustering of such sequences.

Notice, however, that both in supervised and unsupervised sequence classification, each sequence is considered as belonging to a specific class as a whole. A substantially more difficult problem, that is the focus of this project, is that of discovering patterns of sub-sequences in the data. To illustrate this point, Table 2 (left) contains a number of sequences over the alphabet {a,b,c}, a possible set of sub-sequences that might have generated them, and the corresponding segmentation. This is a far more challenging problem than sequence classification and cluster-

sequences	possible segmentation	possible constituents	fixed-length representation								
			ab	ab	ac	ba	bb	bc	ca	cb	cc
acbca	acb|ca	aba, acb, ca	0	0	1	0	0	1	1	0	0
abaca	aba|ca		0	1	1	1	0	0	1	0	0
acbcaabaacb	acb|ca|aba|acb		2	1	2	1	0	1	1	2	0
abaacbaba	aba|acb|aba		1	2	1	3	0	0	0	1	0

Table 2: Left: sequences and their possible constituents. Right: possible fixed-length representation: column xy contains, for each sequence, the # of transitions between symbol x and y.

ing because we need to find an optimal clustering of the sub-sequences at the same time as we find an optimal segmentation of the original sequences. The size of the search space, therefore, increases combinatorially with the cardinality of the alphabet, and, especially, with the length of the sequences, that is normally not bounded[1]. The difficulty of this task is even more evident if we consider noisy and continuous data.

The reason why this specific problem is so central for the future of autonomous systems relying on machine learning is that most practical applications involve continuous streams of unsegmented data. The way these problems have been addressed, so far, is to augment the observations with annotations that greatly reduce the search space, allowing the use of simpler learning methods. However, this practice has at least three drawbacks: i) the large costs of the annotation effort limits the applicability of the methods, ii) the system designer needs to make choices regarding the relevant units to annotate that may not be optimal for the problem at hand, and iii) the ability of the system to keep on learning while deployed is very limited.

[1] however, we could assume a maximum length for the sequences imitating the effect of limited working memory in the brain.

The methods we propose will overcome the above limitations by allowing more flexible learning from unlabelled data. This will reduce the costs of developing Machine Learning-based systems in novel situations. It has the potential of resulting in more efficient models because the fundamental building blocks for the problem will be discovered from the data. Finally, it will allow for life-long learning, where the systems adapt to new situations even after deployment.

In the area of Automatic Speech Recognition (ASR), for example, these new methods will allow fast development of ASR for languages such as Scandinavian languages, that have smaller populations (and consumer base) and therefore smaller amounts of high quality speech resources, compared to "large" languages, such as American English and Mandarin. The methods will allow us to investigate if we can find more optimal fundamental units for ASR than those defined by linguists. They will make it possible to develop dialogue systems (and therefore robotic agents) that learn new words from interacting with their users.

2 Survey of the field

Machine learning methods have traditionally solved the problem of how to efficiently fit the parameters of a statistical model to a particular set of data in order to satisfy some optimality criterion. This is true both for supervised and unsupervised learning problems. In the latter case, although the objective of learning is not explicitly specified in the data, the model structure is usually predetermined. Recently, researchers have stressed the need for more flexible machine learning methods capable of learning some structure of the data, and thus reducing the intervention of the model designer.

The simplest example of this is the inference of the number of classes in clustering. In the past, a vast number of criteria have been proposed to achieve this task[2], In some cases[3], the task has been successfully integrated into the parameter fitting optimality criterion, resulting in closed-form learning rules.

In the attempt to learn more and more of the structure in the data, increasing focus has been recently given to non-parametric Bayesian methods, or infinite latent variable models[4]. These methods let the data determine the number and nature of their parameters more in general than in the clustering example.

2.1 Survey of the field in learning sequences

The models of choice for learning variable length sequences are Markov Chains and Hidden Markov Models[5]. These are graphical models that grasp some of the evolution of the data. For example, simple HMMs are able to distinguish sequences as those in Table 1. Similarly, if the problem is to cluster those sequences in an unsupervised manner, many methods have been proposed. Porikli[6] performs the clustering in the HMM parameters space by Eigenvector Decomposition. Wang and Zhang[7] expand Porikli's ideas by proposing a more principled way

[2]Glenn Milligan and Martha Cooper. "An Examination of Procedures for Determining the Number of Clusters in a Data Set". In: *Psychometrika* 50 (1985), pp. 159–179.

[3]Mario A.T. Figueiredo and Anil K. Jain. "Unsupervised Learning of Finite Mixture Models". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 24.3 (2002), pp. 381–396.

[4]Tom Griffiths and Zoubin Ghahramani. "Infinite latent feature models and the Indian buffet process". In: *NIPS*. 2005.

[5]L. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proc. IEEE* 77.2 (Feb. 1989), pp. 257–286.

[6]Faith Porikli. "Clustering Variable Length Sequences by Eigenvector Decomposition using HMM". in: *Structural, Syntactic, and Statistical Pattern Recognition*. Vol. 3138. LNCS. Springer-Verlag, 2004, pp. 352–360.

[7]Fei Wang and Changshui Zhang. "Pattern Recognition and Data Mining". In: vol. 3686/2005. Springer Berlin / Heidelberg, 2005. Chap. Spectral Clustering for Time Series, pp. 345–354. URL: http://www.

to define the affinity matrix from the HMM parameters, to make two examples.

In the above cases, however, the topology and size of the HMMs have to be predetermined by the experimenter. An attempt to learn the size and topology of the HMM from data was proposed by Beal *et al.*[8]. They introduced an infinite latent variable solution to the HMM inference problem. This new class of models, called infinite HMMs (iHMMs in short), can grow the number of hidden states depending on the data they observe.

However, if we consider the problem illustrated by Table 2 (left), none of the above methods are suitable, because they require a given segmentation of the data. A set of methods that address this problem rely on defining a *fixed-length* representation of each sequence that retains some (local) information about evolution within the sequence. The problem of finding recurrent patterns or sub-sequences, is then shifted to finding structure in the resulting representation. The representation proposed by Stouten *et al.*[9], simply counts the number of transitions between any pairs of symbols in the alphabet (see right side of Table 2 for an example). Then a form of factor analysis based on non-negative matrix factorisation[10] was used to recover the constituent sub-sequences. This paradigm has had a large impact in many areas of Machine Learning. In the context of sequence segmentation, it has shown some success for small number of sub-sequences and short containing sequences. However, its limitations reside in the fact that the representation, such as the one in Table 2 (right), only retains local information about the evolution of the sequences, and long term dependencies are discarded.

Recently, Stepleton[11] proposed a method called Block Diagonal Infinite HMM (BDiHMM) based on the infinite HMM introduced above. During inference, the number of states is allowed to grow indefinitely as in iHMMs, but each state is also assigned to a block. States belonging to the same block are encouraged to have stronger connections and the number of blocks is also inferred from the data. The purpose of the model is to discover sub-behaviours in the data (as in Table 2, left) by means of these blocks. A drawback of these models is that they use sampling methods, such as Monte Carlo Markov Chain (MCMC), for the inference, which is computationally intensive and limits their applicability to real life problems.

Progress beyond the state-of-the-art on learning sequences In this project we intend to advance the state-of-the-art in this area in a number of ways: firstly we will investigate optimisation of the inference proposed in Stepleton *et al.*[12]. We will propose more efficient implementations of the methods as well as experiment with (faster) variational techniques instead of MCMC. We will also investigate the possibility to learn structure in a hierarchical way, by recursively using sub-sequences found in the previous phase as inputs to the next phase.

2.2 Survey of the field in discovering speech units

Current ASR systems are usually based on Hidden Markov Models (HMMs). The language units (words and phonemes) that constitute modelling blocks for the HMMs are predefined by experts in linguistics. This is not guaranteed to result in the most efficient modelling and, for this reason, researchers have long tried to propose methods for extracting the optimal linguistic

springerlink.com/content/3xy2a6cl8r4gf8nq/.

[8] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. "The infinite hidden Markov model". In: *Advances in neural information processing systems*. 2001, pp. 577–584.

[9] Veronique Stouten, Kris Demuynck, and Hugo Van hamme. "Automatically Learning the Units of Speech by Non-negative Matrix Factorisation". In: *Proceedings of Interspeech*. Sept. 2007, pp. 1037–1940.

[10] Daniel D. Lee and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (1999), pp. 788–791.

[11] T. S. Stepleton et al. "The block diagonal infinite hidden Markov model". In: *Proc of AI & Statistics*. 2009, pp. 552–559.

[12] Stepleton et al., see n. 11.

units from the data.

Early attempts were based on discovering sub-word units that could efficiently represent given words. Bimbot *et al.*[13] tried to achieve this by a multigram framework. Svendsen *et al.*[14] proposed a data-driven method for extracting possible pronunciations from the data for a given lexicon. Singh *et al.*[15] extend these ideas to a Large Vocabulary Speech Recognition scenario.

More recently, the field has been revitalised by the attempt to model language acquisition in human infants. A number of methods have been proposed to find linguistic units in a completely unconstrained manner (that is, with no inserted linguistic knowledge). For example Stouten *et al.*[16], already mentioned earlier, use Non-negative Matrix Factorisation and a fixed length representation (see Table 2, right) to discover recurrent patterns in speech. Similarly, Rasanen[17] uses transitional probabilities between atomic acoustic events in order to detect recurring patterns in speech. The method by Park and Glass[18] is based on a dynamic time-warping (DTW) whereas Aimetti[19] proposed the so called DPn-gram method based on Dynamic Programming.

Most of the above methods were shown to work for very small vocabularies and a limited number of speakers, but have failed to scale to real-size problems, to date. This field of research is becoming increasingly popular, indicating the need for such methods. As a consequence, Interspeech 2015, one of the main conferences in Speech Processing, defined a special session entitled: "Zero Resource Speech Technologies: Unsupervised Discovery of Linguistic Units".

Progress beyond the state-of-the-art on speech In this area, we plan to apply the general patter discovery methods we will develop to the speech signal. Compared to the state-of-the-art, we will focus on real size problems. Additionally, we will focus on finding hierarchies of patterns from more specific *phoneme-like* to more general, *word-* and *phrase-like* segments. We will test our models' ability to discover patterns that have linguistic significance, but also the possibility to apply those patterns in automatic speech recognition (ASR). Finally, we will investigate how these methods can be used to update ASR models with new words and expressions in a life-long learning fashion, by interacting with the user.

2.3 Survey of the field on learning gestures

In the field of robotics, great efforts have been recently devoted to creating robots that could coexists and naturally interact with humans in everyday situations. In order for this to happen, it is necessary for the robots to understand human behaviour, and, therefore, human pose, actions,

[13] F. Bimbot et al. "Variable-length sequence modeling: multigrams". In: *IEEE Signal Process. Lett.* 2.6 (1995), pp. 111–113; Sabine Deligne and Frederic Bimbot. "Inference of variable-length linguistic and acoustic units by multigrams". In: *Speech Communication* 23 (1997), pp. 223–241.

[14] Trym Holter and Torbjørn Svendsen. "A comparison of lexicon-building methods for subword-based speech recognisers". In: *TENCON '96. Proceedings. 1996 IEEE TENCON. Digital Signal Processing Applications*. Vol. 1. 1996, pp. 102–106; Trym Holter and Torbjørn Svendsen. "Combined Optimisation of Baseforms and Subword Models for an Hmm Based Speech Recogniser". In: *Proceedings of Fourth International Symposium on Signal Processing and Its Applications (ISSPA)*. vol. 1. Aug. 1996; Trym Holter and Torbjørn Svendsen. "Combined optimisation of baseforms and model parameters in speech recognition based on acoustic subword units". In: *IEEE Workshop on Automatic Speech Recognition and Understanding*. 1997, pp. 199–206.

[15] R. Singh, B. Raj, and R. M. Stern. "Automatic generation of subword units for speech recognition systems". In: *IEEE Trans. Speech Audio Process.* 10.2 (2002), pp. 89–99.

[16] Stouten, Demuynck, and Van hamme, see n. 9, p. 4.

[17] O. Räsänen. "A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events". In: *Cognition* 120.2 (2011), pp. 149–176. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2011.04.001.

[18] Alex S. Park and James R. Glass. "Unsupervised Pattern Discovery in Speech". In: *IEEE Trans. Audio, Speech and Lang. Proc.* 16.1 (2008).

[19] G. Aimetti, R. K. Moore, and L. ten Bosch. "Discovering an Optimal Set of Minimally Contrasting Acoustic Speech Units: A Point of Focus for Whole-Word Pattern Matching". In: *Proc. Interspeech*. 2010, pp. 310–313.

and gestures with a communicative meaning.

As in speech, Hidden Markov Models (HMMs) are a common tool for modelling actions. One of the first such systems was by Yamato et al.[20]. Wilson et al.[21] developed a parameterized gesture recognizer where people's motion was recorded with a Polhemus motion capture system. Starner et al.[22] proposed a system to recognise sign language, where each sign word is associated to an HMM with an ad-hoc structure that fits their data, features are determined with computer vision (users wear colored gloves), and a semantic grammar is used to check the validity of phrases. Alon et al.[23] also addressed the sign language recognition problem, using sophisticated visual motion features and a dynamic programming approach to prune multiple hypotheses. The work by Yang et al.[24] aims at recognising complex actions (e.g. sitting on the floor, jumping) using angles between human body parts as features.

In all the above studies, however, the problem of automatic segmentation of the gestures or actions into constituent units is not addressed. An exception to this is the work by Guerra-Filho and Aloimonos[25]. They study human motions by measuring joint angles, and by defining a language of motion whose primitives are inferred by the value of first and second order derivatives of their measurements. Although very appealing, this approach is only applicable when the measurements are collected with high sampling rates and high signal to noise ratios, as in their case.

Progress beyond the state-of-the-art on gestures In this area, we will apply our methods to a number of datasets available through our collaborations. These include recordings of human motion (collected during the Tomsy project), human-robot communicative gestures (from Poeticon++), and Swedish sign language (from TIVOLI). In all cases, we will advance the state-of-the-art by studying if we can automatically segment the data extracting constituent behavioural units that can later be concatenated into complete gestures.

3 Project description

The project is organised in three parallel tracks that will address related problems incrementally in four temporal phases. The three main tracks are: **a)** development and implementation of general purpose pattern discovery methods, **b)** application to speech recognition, **c)** application to gesture/activity recognition. Track b) and c) depend on track a), but they can start from day one of the project because of the preliminary work done so far. In each case, during the different phases of the project we will move incrementally from solving simpler to more complex problems. At every phase the evaluation will be similar, but based on the complexity of the specific problem addressed in that phase. The different phases of the project are detailed below.

[20]Junji Yamato, Jun Ohya, and Kenichiro Ishii. "Recognizing Human Action in Time-Sequential Images using Hidden Markov Model". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1992, pp. 379–385.

[21]Andrew D. Wilson and Aaron F. Bobick. "Parametric Hidden Markov Models for Gesture Recognition". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 21 (9 1999), pp. 884–900.

[22]Thad Starner, Joshua Weaver, and Alex Pentland. "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (12 1998), pp. 1371–1375.

[23]Jonathan Alon et al. "A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 31.9 (Sept. 2009), pp. 1685–1699.

[24]Hee-Deok Yang, A-Yeon Park, and Seong-Whan Lee. "Gesture Spotting and Recognition for Human–Robot Interaction". In: *IEEE Transactions on Robotics* 23.2 (2007), pp. 256–270.

[25]Gutemberg Guerra-Filho and Yiannis Aloimonos. "A Language for Human Action". In: *Computer* 40.5 (2007), pp. 42–51. ISSN: 0018-9162. DOI: http://doi.ieeecomputersociety.org/10.1109/MC.2007.154.

Phase 1 (1st–14th month) **Track a)**: Based on the already developed algorithm (see Section 5), a number of optimisations will be performed. Firstly we will implement the critical parts of the algorithm in a low level programming language (C), furthermore, we will parallelise the process, and implement the relevant parts in CUDA (NVidia parallel computing architecture). The evaluation will consist in measuring the accuracy and speed of the different versions of the algorithm with synthetic *ad hoc* data. **Track b)**: The implementation of the algorithm already available (Section 5) will be tested on limited speech examples from the TIDIGIT database available at TMH, containing recordings of spoken digits in American English. The tests will start using a limited number of speakers and more speakers will be added as the project evolves and more efficient versions of the algorithm become available. The evaluation will be performed by analysing the obtained patterns, and by using them as building blocks for speech recognition. **Track c)**: We will analyse the classification problems for the available data from Tomsy (human activity), TIVOLI (sign language), and Poeticon++ (communicative gestures). We will assess the difficulty of each problem, in order to select the scenario that is most suitable for the first tests, that will be performed in the following phases of the project. The evaluation will rely on the standard classifiers developed in the corresponding projects in order to score the difficulty of the problems at hand.

Phase 2 (15th–28th month) **Track a)**: we will investigate the possibility of using **variational** algorithms instead of **MCMC** sampling. The evaluation will be carried out in a similar way as in Phase 1. **Track b)**: we will gradually increase the complexity of the problem, by including more speakers and more general speech recordings. First attempts will be made at analysing recordings from the TIMIT speech database available at TMH and containing full sentences. This data set, given its contents, is considered a reliable test bench for speech recognition. Because it is phonetically transcribed, the database will also constitute a great resource for analysing the nature of the patterns discovered by our algorithm. **Track c)**: A first attempt will be made at applying the algorithm to gesture/activity material. Depending on the results from Phase 1, the simplest problem will be addressed first. Evaluation will be based, at this stage, on comparison between the obtained patterns and the sub-gestures defined by the databases.

Phase 3 (29th–42th month) More focus will be given in this phase to the gesture/activity analysis and recognition of **Track c**. The more difficult problems left out in Phase 2, will be addressed here. At the end of this phase we expect the algorithm to be able to extract information that is useful for activity/gesture segmentation and recognition. The evaluation at this point will be comparing the results we obtain with our method to the state-of-the-art of that particular set of data. In **Track b** we will perform full-scale recognition on the TIMIT database. Particular effort will be directed to cleaning up the code from **Track a**, and the resourced developed in **Track b** and **c** for sharing with the community.

Phase 4 (43th–48th month) In this phase we will wrap up the results from the previous phases, complete them if necessary and focus on dissemination. We will extend the conference publications from previous phases in order to publish in journals, and we will advertise the results of the project in all proper venues. We will also take advantage on our national and international collaborations in order to propose the use of our results in future research projects.

4 Significance

The machine learning problem the project aims to solve has a significance for a wide range of fields. It is relevant to all those cases that involve sequences of inputs. It will benefit fields such as speech research, linguistics, computer vision, robotics, econometrics, bioinformatics.

In the speech area, we aim at providing a way to define optimal speech units for speech

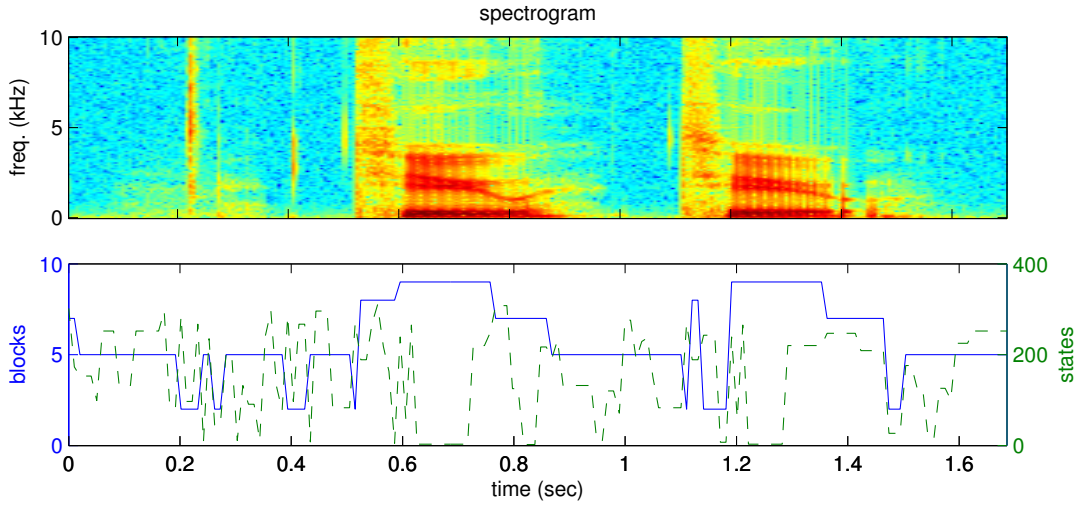

Figure 2: Preliminary results in speech (utterance "two two")

analysis and recognition. We will be able to verify if the assumption of speech being divided into phonemes is optimal in this context, or if better units can be defined. These results have a potential impact on speech recognition in general, regardless the specific language considered in this project. Based on these methods, it will be easier to develop Automatic Speech Recognition systems for languages that lack large quantities of high quality resources. This is the case, for example, of Scandinavian languages that have relatively small populations compared to "large" languages as American English and Mandarin. The methods will also be the basis for more flexible dialogue systems and spoken human-robot interaction. They will allow the user to teach the system new words during the spoken interaction, thus increasing the naturalness of human-machine conversation.

In the area of human activity and gesture recognition, we will provide a way to automatically segment the data, thus proposing a possible solution to the problem of life-long learning. This will provide one of the building blocks to developing systems that are truly autonomous because they can learn independently of the designer intervention.

We will consider the project successful if it will deliver: a) high impact publications, b) a successful PhD education, c) consolidating our national and international collaborations, thus benefiting not only our group, but KTH as a whole. Furthermore, the methods developed in this project will be made available to the research community. Because of the general nature of the sought solutions, we expect the impact of those results to spread beyond the specific areas described above. We also expect the results of this project to live well beyond the time allotted to it, and new applications to arise that will motivate future research.

We will use the available channels in order to advertise the results of the project. The main conferences in speech and robotics will be addressed, as well as pure machine learning conferences. Whenever possible, we will publish our results in the most renown journals in the area. We will also make an effort to advertise the project in popular media, when suitable.

5 Preliminary results

The ambitious goals of this project are possible because of the preliminary work carried out by the PI and his group, and because of the synergies emerging from the PI's ongoing national and international collaborations. Preliminary results are described below.

We have been involved in modelling language learning in infants for a number of years, participating in national projects (Riksbankens Mille, VR BioASU) as well as international EU

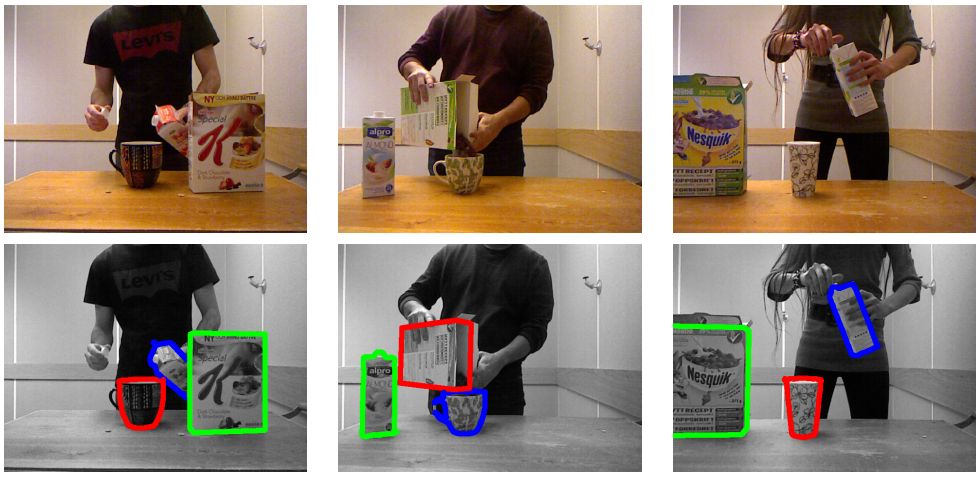

Figure 3: Preliminary results in human activity recognition

projects (Contact, Poeticon++). Some of the results of this research are summarised in Salvi[26]. In particular, we have been studying phenomena such as learning by imitation the mapping between adult speech production and infant (own) speech[27]. We have also been involved in modelling grounding of word meanings in multisensory context in robotic manipulation tasks[28].

More in line with the specific aim of this project, we have introduced a Bayesian non-parametric method for inferring at the same time recurrent patterns and their optimal number in continuous speech[29]. This method, similarly to Stouten *et al.*[30], does not easily scale to medium size vocabularies. We have, therefore, implemented a form of Block Diagonal Infinite HMM[31] and tested it on spoken digit data (see Figure 2). The figure shows that the linguistic units automatically discovered (blocks in the lower graph), are a stable representation for repeating words. This model will constitute the starting point for the work we will carry out in the project.

The PI has also been involved in modelling human gestures in the context of computer vision and robotics, in collaboration with two groups. The first collaboration is with Instituto Superior Técnico in Portugal within the Poeticon++ EU project. Here, we developed a system to classify continuous gestures of a communicative nature[32]. The second collaboration is with the Computer Vision and Active Perception lab within a KTH funded Short Visionary Project and the Tomsy EU project. In this case we created a classifier that uses both computer vision and auditory perception to infer actions in everyday tasks, such as making food[33]. Figure 3 shows a few frames from the database collected in this project that has been made freely available

[26]Giampiero Salvi. "Biologically Inspired Methods for Automatic Speech Understanding". In: *Proc. of BICA*. vol. 196. Advances in Intelligent Systems and Computing. Palermo, Italy, Oct. 2013, pp. 283–286.

[27]Gopal Ananthakrishnan and Giampiero Salvi. "Using Imitation to learn Infant-Adult Acoustic Mappings". In: *Proc. of Interspeech*. Firenze, Italy, 2011.

[28]Giampiero Salvi et al. "Language bootstrapping: Learning word meanings from perception-action association". In: *IEEE Trans. on Syst., Man and Cybern.: Cybern.* 42.3 (June 2012), pp. 660–671.

[29]Niklas Vanhainen and Giampiero Salvi. "Word Discovery with Beta Process Factor Analysis". In: *Proc. of Interspeech*. Portland, OR, USA, Sept. 2012.

[30]Stouten, Demuynck, and Van hamme, see n. 9, p. 4.

[31]Niklas Vanhainen and Giampiero Salvi. "Pattern Discovery in Continuous Speech Using Block Diagonal Infinite HMM". in: *Proc. of IEEE ICASSP*. 2014.

[32]Giovanni Saponaro, Giampiero Salvi, and Alexandre Bernardino. "Robot Anticipation of Human Intentions through Continuous Gesture Recognition". In: *Proc. 4th International Workshop on Collaborative Robots and Human Robot Interaction (CR-HRI)*. San Diego, USA, 2013.

[33]Alessandro Pieropan et al. "Audio-Visual Classification and Detection of Human Manipulation Actions". In: *Proc. of IEEE/RSJ IROS*. 2014.

for the community. The corpora of human activity and gestures recorded during Poeticon++ and Tomsy and the experience gathered, will constitute the basis for the experiments we will perform in the proposed project.

Finally, a substantial number of previous and current studies are relevant to this project because they have the aim of discovering patterns in human-human interaction. For example, we analysed the gaze patterns in a multiparty conversation to infer the degree of involvement of the participants[34]. This study also resulted in a audio-visual corpus that could be used as test for the project[35].

6 International and national collaboration

The project will take place within the School of Computer Science and Communication at KTH, that has a strong background in the areas of speech technology, computer vision, robotics and machine learning. In the latest Research Assessment Exercise (RAE2012), this unit of assessment (Applied Computer Science) was selected as one of the most outstanding at KTH, and the comments were: "Research output is internationally excellent in all fields, with a substantial number of units reaching the level of world-leading quality". In RAE2008, the Speech Group at KTH where the PI is affiliated, was described as: "This is an outstanding, world leading research group — among the top and most respected (a national asset)".

The PI has a number of collaborations within KTH and other European universities, that are key to the success of the project, and are already discussed in Section 5. The international collaboration with Luis Montesano (University of Zaragoza) led to high quality publication in Machine Learning and Cognitive System[36]. Also highly cited was a collaboration with Björn Lindblom (SU) and R. Diehl (U. of Texas) on the emergence of phonological structure in language[37]. Additionally, the PI has active collaborations with Gothenburg Center for Language Technology (CLT) that led to the creation of free acoustic models for automatic speech recognition (ASR)[38] and to a freely available plugin for ASR in the popular software Wavesurfer[39]. Furthermore, the PI's collaboration with the Computer Vision and Active Perception lab at KTH is not limited to the Tomsy project (Hedvig Kjellström), but also includes a Short Visionary Project with the aim of developing new features for audio processing based on Topological Theory, together with Florian Pokorny. The PI was also involved in brain related research together with the Computational Biology group at KTH (Anders Lansner, Pavel Herman), that resulted in a study on speech representations in the brain[40].

The Speech Group where the PI is affiliated, has also been involved with activities that are relevant to the current project. The TIVOLI project (PTS) has collected a database of Swedish sign language that will be used as one of the potential test bases for our methods.

[34]Catharine Oertel and Giampiero Salvi. "A Gaze-based Method for Relating Group Involvement to Individual Engagement in Multimodal Multiparty Dialogue". In: *Proc. of the ACM International Conference on Multimodal Interaction (ICMI)*. Sydney, Australia, 2013.

[35]Catharine Oertel et al. "The KTH Games Corpora: How to Catch a Werewolf". In: *IVA 2013 Workshop Multimodal Corpora: Beyond Audio and Video - MMC*. Edinburgh, UK, 2013.

[36]Salvi et al., see n. 28, p. 9.

[37]Björn Lindblom et al. "Sound systems are shaped by their users: The recombination of phonetic substance". In: *Where do features come from? Cognitive, physical and developmental bases of distinctive speech categories*. Ed. by G. Nick Clements and Rachid Ridouane. Vol. 6. John Benjamins Publishing Company, 2011, pp. 67–98.

[38]Niklas Vanhainen and Giampiero Salvi. "Free Acoustic Models for Large Vocabulary Continuous Speech Recognition in Swedish". In: *Proc. of the Language Resources and Evaluation Conference (LREC)*. 2014.

[39]Giampiero Salvi and Niklas Vanhainen. "The WaveSurfer Automatic Speech Recognition Plugin". In: *Proc. of the Language Resources and Evaluation Conference (LREC)*. 2014.

[40]Tin Franovic et al. "Cortex-inspired network architecture for large-scale temporal information processing". In: *Frontiers in neuroinformatics*. 2013.

Interdisciplinarity

My application is interdisciplinary

☐

An interdisciplinary research project is defined in this call for proposals as a project that can not be completed without knowledge, methods, terminology, data and researchers from more than one of the Swedish Research Councils subject areas; Medicine and health, Natural and engineering sciences, Humanities and social sciences and Educational sciences. If your research project is interdisciplinary according to this definition, you indicate and explain this here.

Click here for more information

Scientific report

Scientific report/Account for scientific activities of previous project

---

Budget and research resources

Project staff

Describe the staff that will be working in the project and the salary that is applied for in the project budget. Enter the full amount, not in thousands SEK.

Participating researchers that accept an invitation to participate in the application will be displayed automatically under Dedicated time for this project. Note that it will take a few minutes before the information is updated, and that it might be necessary for the project leader to close and reopen the form.

Dedicated time for this project

	Role in the project	Name	Percent of full time
1	Applicant	Giampiero Salvi	

Salaries including social fees

	Role in the project	Name	Percent of salary	2016	2017	2018	2019	Total
1	Applicant	Giampiero Salvi	30	260,000	267,000	273,000	280,000	1,080,000
2	Other personnel without doctoral degree	N/N doktorand	100	534,000	587,000	646,000	711,000	2,478,000
	Total			794,000	854,000	919,000	991,000	3,558,000

Other costs

Describe the other project costs for which you apply from the Swedish Research Council. Enter the full amount, not in thousands SEK.

Premises

	Type of premises	2016	2017	2018	2019	Total
1	kontorslokal	96,000	103,000	111,000	120,000	430,000
	Total	96,000	103,000	111,000	120,000	430,000

Running Costs

	Running Cost	Description	2016	2017	2018	2019	Total
1	Resor	konferenser, workshop	30,000	30,000	30,000	30,000	120,000
2	Dator	2 st	60,000				60,000
3	Publisering		15,000	15,000	15,000	15,000	60,000
	Total		105,000	45,000	45,000	45,000	240,000

Depreciation costs

	Depreciation cost	Description	2016	2017	2018	2019

Total project cost

Below you can see a summary of the costs in your budget, which are the costs that you apply for from the Swedish Research Council. Indirect costs are entered separately into the table.

Under Other costs you can enter which costs, aside from the ones you apply for from the Swedish Research Council, that the project includes. Add the full amounts, not in thousands of SEK.

The subtotal plus indirect costs are the total per year that you apply for.

Total budget

Specified costs	2016	2017	2018	2019	Total, applied	Other costs	Total cost
Salaries including social fees	794,000	854,000	919,000	991,000	3,558,000		3,558,000
Running costs	105,000	45,000	45,000	45,000	240,000		240,000
Depreciation costs					0		0
Premises	96,000	103,000	111,000	120,000	430,000		430,000
Subtotal	995,000	1,002,000	1,075,000	1,156,000	4,228,000	0	4,228,000
Indirect costs	412,000	443,000	477,000	515,000	1,847,000		1,847,000
Total project cost	1,407,000	1,445,000	1,552,000	1,671,000	6,075,000	0	6,075,000

Explanation of the proposed budget

Briefly justify each proposed cost in the stated budget.

Explanation of the proposed budget*

Salaries

The requested grant is meant to cover salaries and equipment for the PI (30%), and a doctoral student (100%) that will be employed specifically for the project.

Travel

We included **30kSEK/year** for travelling to conferences, considering that some of those costs for the PhD student may be covered by student travel grants.

Equipment

We included in the budget **60 kSEK** for a laptop for the new doctoral student and for a fast computer that will be used for simulations, together with the other machines available at CSC.

Publications

We will try to publish on free open access journals every time this is possible. For the other cases, we added **15 kSEK/year** of publication costs based on the IEEE open access publication fee. We considered one such publication per yea

Other funding

Describe your other project funding for the project period (applied for or granted) aside from that which you apply for from the Swedish Research Council. Write the whole sum, not thousands of SEK.

Other funding for this project

Funder	Applicant/project leader	Type of grant	Reg no or equiv.	2016	2017	2018	2019

CV

Giampiero Salvi

1 Higher education qualification:

1998 MSc. Electrical Engineering, Università La Sapienza, Rome, Italy

2 Degree of Doctor:

2006 PhD. Computer Science, KTH, Stockholm, Sweden
Title: “Mining Speech Sounds”, supervisor: Professor Björn Granström

3 Postdoctoral positions:

2007–2009 Instituto Superior Técnico, Inst. for Systems and Robotics, Lisbon, Portugal

4 Docent level

2012 Docent in Computer Science, Speech Communication, KTH, Stockholm, Sweden

5 Present position:

2013-01-01 Associate Professor in Machine Learning, KTH. 50% research.

6 Previous employments

2010-01-01–2012-12-31	Assistant Professor, KTH, CSC, TMH, Stockholm, Sweden
2009-07-01–2009-12-31	Researcher, KTH, TMH, Stockholm, Sweden
2007-03-01–2009-06-01	Researcher and Postdoc, Instituto Superior Técnico, Institute for Systems and Robotics, Lisbon, Portugal
2006–2007	Researcher, KTH, TMH, Stockholm, Sweden
2002–2006	PhD Student, KTH, TMH, Stockholm, Sweden
1999–2001	Research Assistant, KTH, TMH, Stockholm, Sweden

7 PhD supervision:

Niklas Vanhainen, main supervisor, ongoing
Alessandro Pieropan (CVAP, KTH), assistant supervisor, ongoing
Giovanni Saponaro (IST, Portugal), assistant supervisor, ongoing
Gopal Ananthakrishnan, assistant supervisor, defended on 2012-02-27
Chris Koniaris, assistant supervisor, defended 2012-10-05
Daniel Neiberg, assistant supervisor, defended 2012-09-28

8 Additional information of relevance to the application

GS has a long experience in several areas of applied Machine Learning. His main area of investigation is Speech Technology, but he has also worked in close collaboration with highly renown groups in Computer Vision, Cognitive Systems and Robotics, both nationally and internationally. He has always published in flagship conferences (Interspeech, IEEE ICASSP, IEEE ICRA, IEEE IROS, ...) and journals (JASA, Speech Communication, IEEE Systems, Man, and Cybernetics, ...) for such areas. GS main interests are in unsupervised Machine Learning methods and algorithms. His most cited paper is on devising standard procedures for building speech recognition across languages (RefRec). This in collaboration with Norwegian, Danish and Slovenian universities and companies. His second most cited paper is about the Synface system where GS developed the core methods and technology for low latency phonetic recognition that were later commercialised by the SynFace AB company, of which he is co-founder. In this work GS closely collaborated with University College London and Viataal The Netherlands. GS’s collaboration with Björn Lindblom (SU) and R. Diehl (U. of Texas), led to another highly cited work on the origin of phonological structure in language. Finally the long lasting collaboration with J.A. Santos-Victor and A. Bernardino (IST, Portugal) and L. Montesano (U. of Zaragoza, Spain), led, among others, to a publication in IEEE SMC that is one of the highest impact journal in the field of cognitive systems. This work is particularly relevant to the project because it is an attempt to model robotic language acquisition in similar multimodal settings as in humans. The collaboration with Piero Cosi and colleagues (CNR, Padova, Italy), led to the discovery of patterns related to emotional content in speech production. Other national collab-

orations that are relevant to the project are those with Hedvig Kjellström (CVAP, KTH) that led to the data collection on human activities that will be used by part of the project, and Florian Pokorny (CVAP, KTH) which is ongoing and has the high ambition to introduce novel features for speech analysis and recognition by means of advanced Topological Theory.

Personal Grants

2007–2009	Portuguese Research Council (Fundação para Ciência e a Tecnologia, FCT), postdoctoral grant, Portugal (3 years)
2010–2013	Vetenskapsrådet, forskarassistent, "Biologically Inspired Statistical Methods for Automatic Speech Understanding" (4+1 years)

Review / referee assignments in international periodicals

GS is regularly reviewing for the following international conferences: Interspeech, IEEE ICRA, IEEE IROS, IEEE ICDL, AVSP, NODALIDA) and for the following journals: Pattern Recognition (Elsevier), Acta Cybernetica, Journal of Logopedics Phoniatrics Vocology (Psychology Press), IEEE Transactions on Systems, Man, and Cybernetics, part B., IEEE Transactions on Circuits and Systems for Video Technology, Speech Communication (Springer), Signal, Image and Video Processing Journal (Springer), Journal on Computer Speech and Language (Elsevier).

Conference and workshop committee

2011	Audio-Visual Speech Processing (AVSP), organising committee, Volterra, Italy

Assignments as public examiner/opponent at PhD defence (KTH)

2012	Benjamin Auffarth, KTH/CSC/CB, Sweden, **commission member**
2012	Samer Al Moubayed, KTH/CSC/TMH, Sweden, **internal review**
2012	Yasemin Bekiroglu, KTH/CSC/CVAP, Sweden, **commission member**
2013	Gael Dubus, KTH/CSC/TMH, Sweden, **internal review**
2013	Gustav Henter, KTH/EE/KT, Sweden, **commission member**

Assignments as public examiner/opponent at PhD defence (External)

2012	Jonas Hornstein, IST, Portugal, **first opponent**
2013	Jarle Bauck Hamar, NTNU, Norway, **second opponent**

Assignments as outside expert

2006	the Wiener Wissenschafts, Forschungs- und Technologiefonds (Vienna Science and Technology Fund) Five Senses project call. The fund delivered 15M euros to 9 of the 34 submitted projects
2011	the United States–Israel Binational Science Foundation, 2011 call for project proposals in the area of Computer Sciences: Artificial Intelligence, Natural Language Processing, Machine Learning
2013	the United States–Israel Binational Science Foundation, 2013 call for project proposals in the area of Computer Sciences: Artificial Intelligence, Natural Language Processing, Machine Learning

Exibitions and Dissemination

Sep 2001	showcased the Teleface project at the i3 Research Village at Comdex in Basel, Switzerland
Jun 2006	invited by the European Commission to show the results of the Synface project at the Riga Ministerial Conference "ICT for an Inclusive Society"
Nov 2012	invited lecture to Ericsson Innovation Day
Nov 2013	invited lecture on Automatic Speech Recognition at Hennes&Mauritz AB
Dec 2013	invited participant to Crosstalks Talk Show: "The power of the human voice" `http://crosstalks.tv/talks/the-power-of-the-human-voice/`

Publication list

Publications since 2007

The following sections list publications coauthored by the PI since 2007. Citations are according to Google Scholar last updated on 2015-03-06. Google Scholar reports a total of 464 citations, an h-index of 12 and a i10-index of 14.

The five publications of special interest for the proposal are: [J-4, C-2, C-6, C-14, B-1]

1 Peer reviewed original articles:

[J-1] Sofia Strömbergsson, **Giampiero Salvi**, and David House. "Acoustic and perceptual evaluation of category goodness of /t/ and /k/ in typical and misarticulated children's speech". In: *Journal of the Acoustical Society of America* (in press) (2015).
Number of citations: –

[J-2] Christos Koniaris, **Giampiero Salvi**, and Olov Engwall. "On Mispronunciation Analysis of Individual Foreign Speakers Using Auditory Periphery Models". In: *Speech Communication* 55.5 (2013), pp. 691–706.
Number of citations: 1

[J-3] Daniel Neiberg, **Giampiero Salvi**, and Joakim Gustafson. "Semi-supervised methods for exploring the acoustics of non-lexical feedback in Swedish". In: *Speech Communication* 55.3 (2013), pp. 451–469.
Number of citations: 5

[J-4] **Giampiero Salvi**, Luis Montesano, Alexandre Bernardino, and José Santos-Victor. "Language bootstrapping: Learning word meanings from perception-action association". In: *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics* 42.3 (June 2012), pp. 660–671.
Number of citations: 14

[J-5] **Giampiero Salvi**, Jonas Beskow, Samer Al Moubayed, and Björn Granström. "SynFace — Speech-driven Facial Animation for Virtual Speech-reading Support". In: *EURASIP Journal on Audio, Speech, and Music Processing* (Sept. 2009).
Number of citations: 17

2 Peer reviewed conference contributions:

[C-1] Alessandro Pieropan, **Giampiero Salvi**, Karl Pauwels, and Hedvig Kjellström. "A dataset of human manipulation actions". In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 2014.
Number of citations: –

[C-2] Alessandro Pieropan, **Giampiero Salvi**, Karl Pauwels, and Hedvig Kjellström. "Audio-Visual Classification and Detection of Human Manipulation Actions". In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2014.
Number of citations: 2

[C-3] **Giampiero Salvi** and Niklas Vanhainen. "The WaveSurfer Automatic Speech Recognition Plugin". In: *Proc. of the Language Resources and Evaluation Conference (LREC)*. 2014.
Number of citations: 1

[C-4] Sofia Strömbergsson, **Giampiero Salvi**, and David House. "Gradient evaluation of /k/-likeness in typical and misarticulated child speech". In: *Proceedings of the International Clinical Phonetics and Linguistics Association (ICPLA) Conference*. 2014.
Number of citations: –

[C-5] Niklas Vanhainen and **Giampiero Salvi**. “Free Acoustic Models for Large Vocabulary Continuous Speech Recognition in Swedish”. In: *Proc. of the Language Resources and Evaluation Conference (LREC)*. 2014.
Number of citations: 1

[C-6] Niklas Vanhainen and **Giampiero Salvi**. “Pattern Discovery in Continuous Speech Using Block Diagonal Infinite HMM”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014.
Number of citations: –

[C-7] Tin Franovic, Pavel Herman, **Giampiero Salvi**, Simon Benjaminsson, and Anders Lansner. “Cortex-inspired network architecture for large-scale temporal information processing”. In: *Frontiers in neuroinformatics*. 2013.
Number of citations: –

[C-8] Catharine Oertel and **Giampiero Salvi**. “A Gaze-based Method for Relating Group Involvement to Individual Engagement in Multimodal Multiparty Dialogue”. In: *Proc. of the ACM International Conference on Multimodal Interaction (ICMI)*. Sydney, Australia, 2013.
Number of citations: 3

[C-9] Catharine Oertel, **Giampiero Salvi**, Jana Götze, Jens Edlund, Joakim Gustafson, and Mattias Heldner. “The KTH Games Corpora: How to Catch a Werewolf”. In: *IVA 2013 Workshop Multimodal Corpora: Beyond Audio and Video - MMC*. Edinburgh, UK, 2013.
Number of citations: 3

[C-10] **Giampiero Salvi**. “Biologically Inspired Methods for Automatic Speech Understanding”. In: *Proceedings of the Annual International Conference on Biologically Inspired Cognitive Architectures (BICA)*. Vol. 196. Advances in Intelligent Systems and Computing. Palermo, Italy, Oct. 2013, pp. 283–286.
Number of citations: –

[C-11] Giovanni Saponaro, **Giampiero Salvi**, and Alexandre Bernardino. “Robot Anticipation of Human Intentions through Continuous Gesture Recognition”. In: *Proc. 4th International Workshop on Collaborative Robots and Human Robot Interaction (CR-HRI)*. San Diego, USA, 2013.
Number of citations: 3

[C-12] Christos Koniaris, Olov Engwall, and **Giampiero Salvi**. “Auditory and Dynamic Modeling Paradigms to Detect L2 Mispronunciations”. In: *Proceedings of International Conference on Speech Communication and Technology (Interspeech)*. Portland, OR, USA, Sept. 2012.
Number of citations: 1

[C-13] Christos Koniaris, Olov Engwall, and **Giampiero Salvi**. “On the Benefit of Using Auditory Modeling for Diagnostic Evaluation of Pronunciations”. In: *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*. Stockholm, Sweden, 2012, pp. 59–64.
Number of citations: 2

[C-14] Niklas Vanhainen and **Giampiero Salvi**. “Word Discovery with Beta Process Factor Analysis”. In: *Proceedings of International Conference on Speech Communication and Technology (Interspeech)*. Portland, OR, USA, Sept. 2012.
Number of citations: 4

[C-15] Gopal Ananthakrishnan and **Giampiero Salvi**. “Using Imitation to learn Infant-Adult Acoustic Mappings”. In: *Proceedings of International Conference on Speech Communication and Technology (Interspeech)*. Firenze, Italy, 2011.
Number of citations: 5

[C-16] **Giampiero Salvi**, Fabio Tesser, Enrico Zovato, and Piero Cosi. “Analisi Gerarchica degli Inviluppui Spettrali Differenziali di una Voce Emotiva”. In: *Proceedings of AISV 2011*. Lecce, Italy, 2011.
Number of citations: –

[C-17] **Giampiero Salvi**, Fabio Tesser, Enrico Zovato, and Piero Cosi. “Cluster Analysis of Differential Spectral Envelopes on Emotional Speech”. In: *Proceedings of International Conference on Speech Communication and Technology (Interspeech)*. Makuhari, Japan, 2010.
Number of citations: 4

[C-18] Samer Al Moubayed, Jonas Beskow, Ann-Marie Öster, **Giampiero Salvi**, Björn Granström, N. van Son, E. Ormel, and T. Herzke. “Virtual Speech Reading Support for Hard of Hearing in a Domestic Multi-media Setting”. In: *Proceedings of International Conference on Speech Communication and Technology (Interspeech)*. Brighton UK, 2009.
Number of citations: 2

[C-19] Jonas Beskow, **Giampiero Salvi**, and Samer Al Moubayed. “Synface: Verbal and Non-verbal Face Animation from Audio”. In: *Proceedings of Audio-Visual Speech Processing (AVSP)*. Norwich, England, 2009.
Number of citations: 1

[C-20] Verica Krunic, **Giampiero Salvi**, Alexandre Bernardino, Luis Montesano, and José Santos-Victor. “Affordance based word-to-meaning association”. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Kobe, Japan, 2009.
Number of citations: 14

[C-21] Samer Al Moubayed, Jonas, Beskow, and **Giampiero Salvi**. “SynFace Phone Recognizer for Swedish Wideband and Narrowband Speech”. In: *Proceedings of The second Swedish Language Technology Conference (SLTC)*. Stockholm, Sweden., Nov. 2008.
Number of citations: –

[C-22] Jonas Beskow, Björn Granström, Peter Nordqvist, Samer Al Moubayed, **Giampiero Salvi**, Tobias Herzke, and Arne Shultz. “Hearing at Home – Communication support in home environments for hearing impaired persons”. In: *Proceedings of International Conference on Speech Communication and Technology (Interspeech)*. Brisbane, Australia, Sept. 2008.
Number of citations: 7

[C-23] Verica Krunic, **Giampiero Salvi**, Alexandre Bernardino, Luis Montesano, and José Santos-Victor. “Associating word descriptions to learned manipulation task models”. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. Nice, France, 2008.
Number of citations: 2

3 Book chapters:

[B-1] Björn Lindblom, R. Diehl, S.-H. Park, and **Giampiero Salvi**. “Sound systems are shaped by their users: The recombination of phonetic substance”. In: *Where do features come from? Cognitive, physical and developmental bases of distinctive speech categories*. Ed. by G. Nick Clements and Rachid Ridouane. Vol. 6. John Benjamins Publishing Company, 2011, pp. 67–98.
Number of citations: 39

4 Other publications:

[O-1] **Giampiero Salvi** and Samer Al Moubayed. “Spoken Language Identification using Frame Based Entropy Measures”. In: *Proceedings of Fonetik*. Stockholm, Sweden, 2011.
Number of citations: –

[O-2] Samer Al Moubayed, Jonas Beskow, Ann-Marie Öster, **Giampiero Salvi**, Björn Granström, N. van Son, E. Ormel, and T. Herzke. “Studies on Using the SynFace Talking Head for the Hearing Impaired”. In: *Proceedings of Fonetik*. 2009.
Number of citations: 3

[O-3] Björn Lindblom, Randy Diehl, Sang-Hoon Park, and **Giampiero Salvi**. “(Re-)use of place features in voiced stop inventories: Role of articulatory, perceptual and developmental constraints”. In: *Proceedings of Fonetik*. University of Gothenburg, June 2008, pp. 5–8.
Number of citations: 2

5 Open access computer programs that you have developed

- active development in the Wavesurfer speech analysis software (main developers Jonas Beskow and Kåre Sjölander)

  `http://sourceforge.net/projects/wavesurfer/`

- The Snack Sndfile Extention, In collaboration with Giulio Paci

  `https://github.com/snacksndfileext`

- The Wavesurfer Automatic Speech Recognition Plugin

  `http://www.speech.kth.se/asr/`

- Free Acoustic and Language models for ASR

  `http://www.speech.kth.se/asr/`

6 Publications with highest citation count:

The following are the publications with the 5 highest citation counts regardless the year of publication.

- Borge Lindberg, Finn Tore Johansen, Narada Warakagoda, Gunnar Lehtinen, Zdravko Kačič, Andrei Žgank, Kjell Elenius, and **Giampiero Salvi**. “A noise robust multilingual reference recogniser based on SpeechDat(II)”. in: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. 2000
  Number of citations: 82

- Jonas Beskow, Inger Karlsson, Joe Kewley, and **Giampiero Salvi**. “SYNFACE - A Talking Head Telephone for the Hearing-impaired”. In: *Computers Helping People with Special Needs*. Ed. by K. Miesenberger, J. Klaus, W. Zagler, and D. Burger. Vol. 3118. Lecture Notes in Computer Science. Springer-Verlag, 2004, pp. 1178–1186
  Number of citations: 59

- Finn Tore Johansen, Narada Warakagoda, Borge Lindberg, Gunnar Lehtinen, Zdravko Kačič, Andrei Žgank, Kjell Elenius, and **Giampiero Salvi**. “The COST 249 SpeechDat multilingual reference recogniser”. In: *Proceedings of XLDB Workshop on Very Large*

Telephone Speech Databases. Athens, Greece, 2000
Number of citations: 43

- Björn Lindblom, R. Diehl, S.-H. Park, and **Giampiero Salvi**. “Sound systems are shaped by their users: The recombination of phonetic substance”. In: *Where do features come from? Cognitive, physical and developmental bases of distinctive speech categories*. Ed. by G. Nick Clements and Rachid Ridouane. Vol. 6. John Benjamins Publishing Company, 2011, pp. 67–98
  Number of citations: 39
- Inger Karlsson, Andrew Faulkner, and **Giampiero Salvi**. “SYNFACE - a talking face telephone”. In: *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*. 2003, pp. 1297–1300
  Number of citations: 18

CV

Name: Giampiero Salvi	**Doctorial degree:** 2006-11-02
Birthdate: 19730501	**Academic title:** Docent
Gender: Male	**Employer:** Kungliga Tekniska högskolan

Research education

Dissertation title (swe)

Dissertation title (en)

Mining speech sounds : machine learning methods for automatic speech recognition and analysis

Organisation	Unit	Supervisor
Kungliga Tekniska Högskolan, Sweden Sweden - Higher education Institutes	TMH, Tal, musik och hörsel	Björn Granström

Subject doctors degree	ISSN/ISBN-number	Date doctoral exam
10208. Språkteknologi (språkvetenskaplig databehandling)	91-7178-446-2	2006-11-02

Publications

Name: Giampiero Salvi	**Doctorial degree:** 2006-11-02
Birthdate: 19730501	**Academic title:** Docent
Gender: Male	**Employer:** Kungliga Tekniska högskolan

Salvi, Giampiero has not added any publications to the application.

Register

Terms and conditions

The application must be signed by the applicant as well as the authorised representative of the administrating organisation. The representative is normally the department head of the institution where the research is to be conducted, but may in some instances be e.g. the vice-chancellor. This is specified in the call for proposals.

The signature*from the applicant*confirms that:

- the information in the application is correct and according to the instructions form the Swedish Research Council
- any additional professional activities or commercial ties have been reported to the administrating organisation, and that no conflicts have arisen that would conflict with good research practice
- that the necessary permits and approvals are in place at the start of the project e.g. regarding ethical review.

The signature *from the administrating organisation* confirms that:

- the research, employment and equipment indicated will be accommodated in the institution during the time, and to the extent, described in the application
- the institution approves the cost-estimate in the application
- the research is conducted according to Swedish legislation.

The above-mentioned points must have been discussed between the parties before the representative of the administrating organisation approves and signs the application.

Project out lines are not signed by the administrating organisation. The administrating organisation only sign the application if the project outline is accepted for step two.

Applications with an organisation as applicant is automatically signed when the application is registered.