SEMANTIC ANALYSIS OF SPOKEN INPUT USING MARKOV LOGIC NETWORKS

Vladimir Despotovic[1], *Oliver Walter*[2], *Reinhold Haeb-Umbach*[2]

[1]University of Belgrade, Technical Faculty of Bor, Serbia
[2]University of Paderborn, Department of Communications Engineering, Germany
vdespotovic@tf.bor.ac.rs,{walter,haeb}@nt.uni-paderborn.de

Abstract

We present a semantic analysis technique for spoken input using Markov Logic Networks (MLNs). MLNs combine graphical models with first-order logic. They are particularly suitable for providing inference in the presence of inconsistent and incomplete data, which are typical of an automatic speech recognizer's (ASR) output in the presence of degraded speech. The target application is a speech interface to a home automation system to be operated by people with speech impairments, where the ASR output is particularly noisy. In order to cater for dysarthric speech with non-canonical phoneme realizations, acoustic representations of the input speech are learned in an unsupervised fashion. While training data transcripts are not required for the acoustic model training, the MLN training requires supervision, however, at a rather loose and abstract level. Results on two databases, one of them for dysarthric speech, show that MLN-based semantic analysis clearly outperforms baseline approaches employing non-negative matrix factorization, multinomial naive Bayes models, or support vector machines.

Index Terms: Unsupervised learning, Acoustic units, Speech, Markov Logic Networks, Semantic frame

1. Introduction

Semantic analysis is the task of learning a mapping from spoken or written language to a semantic representation, and thus discovering the meaning of an utterance. One of the approaches that are largely used in natural language processing (NLP) to represent meanings is based on semantic frames. Semantic frames are composed of slots, which represent specific attributes of the spoken utterance. The task here is three-fold [1]: i) target word detection finds semantically relevant words in an utterance; ii) frame classification classifies the input utterance into frames that correspond to an action or a domain of interest; and iii) slot filling finds the slot values that correspond to frame attributes of the input utterance [2].

While semantic analysis in NLP assumes processing of typed input (written language), we are interested in determining the meaning of a spoken utterance, where we also have to deal with the inaccuracies of acoustic recognition. A straightforward way to solve this problem is to use an automatic speech recognizer (ASR) that transforms spoken input into word sequences and then apply the techniques already developed for processing written language. However, spoken language is rather spontaneous and often doesn't follow the grammar of a language. Moreover, ASR will inevitably introduce recognition errors. Therefore it is necessary to adapt the natural language semantic analyser to cope with the problems of spoken language. An excellent survey of techniques for the integration of ASR and spoken language understanding can be found in [3].

For the target application considered here, a speech interface to a home automation system for speech-impaired users, off-the-shelf speaker-independent ASR is known to fail, because pronuncations of dysarthric speech deviate from the standard [4]. One approach that proved successful and that avoided the labeling of the speaker-dependent training data and the necessity of a pronunciation lexicon is the "self-learning vocal interface" described in [5, 6]: In the training session the user can choose the words freely by which an action of a home automation system is to be evoked. Only weak supervision is required, i.e., a label which encodes the semantics of an utterance, while no literal transcription of the user's utterance is required. The spoken input is represented by Gaussian posteriorgrams and Non-negative Matrix Factorization (NMF) is used to learn the mapping of utterances to actions.

To keep these benefits we will also bypass word recognition here. The input speech is decoded into a subword unit sequence, where the subword unit models are learnt in an unsupervised fashion as decribed in [7]. However, we developed quite a different approach to map the subword unit sequence to semantics. We propose to use Markov Logic Networks (MLNs), which are a natural choice, since the mapping rules can easily be represented using first-order logic [8]. MLNs have recently been used in different areas of NLP, such as semantic role labelling [9, 10], word sense disambiguation [11], natural language understanding [12, 13] and unsupervised semantic parsing [14]. Contrary to previous work which was done on clean text transcriptions, we propose to use MLNs for the task of semantic analysis on noisy input, i.e. mappings to meaning representations directly from spoken utterances. Although Kennington and Schlangen in [15] use MLNs for situated incremental natural language understanding from the noisy input coming from the output of the ASR, beside the speech they use additional information in the form of discourse context (previous action) and situational context (the current state of the game for the Pentomino game domain), which requires additional annotations. We use only the acoustic representations obtained from the raw speech.

We perform experiments in two domains, a home automation task for speech impaired people (DOMOTICA 3) and a vocally guided card game *patience* (PATCOR) [16].

The remainder of the paper is organized as follows. Section 2 gives a brief overview of the MLN framework. Section 3 describes the acoustic representation of the speech. The speech corpora are presented in Section 4. Section 5 describes in detail the experimental setup, followed by results and discussion in Section 6 and concluding remarks in Section 7.

2. Markov Logic Networks

Markov Logic Networks are a statistical relational learning technique that combines undirected graphical models (Markov networks) and logical reasoning (First-order logic). First-order logic (FOL) formulae are used to define the relations between task elements. By attaching weights to the FOL formulae these relations can be transformed into Markov networks to create statistical models of the task [17]. The probability distribution over a set of random variables $X = (X_1, X_2, ..., X_n)$ that correspond to the groundings of the predicates in FOL formulae is given as:

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_{i=1}^{F} \omega_i \sum_{g \in G_i} g(x)\right), \qquad (1)$$

where F is the total number of FOL formulae, ω_i are weights, G_i are groundings of the FOL formulae f_i and $g(x)$ is a binary function that takes the value 1 if the grounding of the FOL formula is true and 0 otherwise. Hence, the sum $\sum_{g \in G_i} g(x)$ simply counts the true groundings of f_i given the current truth assignment of X, where grounding refers to assigning constants to variables. Z is a normalizing term obtained by forcing the probabilities to sum up to unity. Weights are typically learned from training data. We used the Alchemy 2.0 engine [18] for learning the weights discriminatively using the rescaled conjugate gradient algorithm, while the inference in MLN was performed using the MC-SAT algorithm [19].

3. Acoustic Representation

In order to learn the mappings to semantic representations directly from the raw speech, we employ an intermediate acoustic representation of the spoken input in terms of subword units, called Acoustic Unit Descriptors (AUDs). AUDs have been originally developed for the semantic analysis of multimedia content [20, 21], but then were adapted to the unsupervised learning of speech representations. Each AUD is modeled by a 3-state left-to-right Hidden Markov Model (HMM) with Gaussian mixture output densities. HMM training of models for the AUDs is done in a completely unsupervised fashion. They are similar to the self-organizing units described in [22].

It has been shown in [23] that AUDs are able to capture acoustically consistent phenomena and represent recurring patterns of feature vectors, and furthermore that they are competitive to other unsupervised acoustic learning techniques. For more details about the learning of the acoustic representation the reader is referred to [24].

4. Datasets

For our experiments, we used task-oriented conversational data from the DOMOTICA 3 home automation domain and the PATCOR card game domain, collected in the framework of the ALADIN project [25].

4.1. DOMOTICA 3

The DOMOTICA 3 speech corpus contains recordings of dysarthric speakers controlling a home automation system. The language of the corpus is Belgian Dutch. The corpus was collected in a Wizard-of-Oz study, where the subjects were asked to command 26 distinct actions for the home automation system, which was simulated in a 3D computer animation to ensure an unbiased choice of words and grammar by the user [16].

The total length of the dataset used in our experiments is approximately 4 hours of speech, with 2055 utterances spoken by 9 speakers, 228 per speaker on average. According to speech intelligibility scores obtained using an automated tool [26], all except two speakers were considered to utter dysarthric speech.

A typical command in DOMOTICA 3 is: *ALADIN lichten in de woonkamer en keuken uit* (*ALADIN turn off the lights in the living room and kitchen*). While the commands are fairly short, the major challenge of the dataset is the fact that pronunciation of dysarthric speakers deviates from the non-impared ones: rate of speech is lower, segments are pronounced differently, pronunciation is less consistent [4].

4.2. PATCOR

The PATCOR speech corpus contains recordings of non-pathological, normal speaking subjects playing a vocally guided card game patience (solitaire). The language of the corpus is Belgian Dutch. The average number of moves per game session is 55 [16]. The total length of the dataset is approximately 3 hours and 20 minutes, with 1912 utterances spoken by eight speakers, 239 per speaker on average.

A typical command in PATCOR is: *De harten boer op de klaveren dame* (*Put the Jack of hearts on the Queen of clubs*). Note the importance of the order of words here, where the change of word order would change the meaning of the utterance. Note also that commands such as: *De zwarte dame naar de rode heer* (*Put the black Queen on the red King*) are present in the dataset, where clearly ambigious mappings are possible for both the *black Queen* (spades or clubs) and the *red King* (hearts or diamonds). An additional challenge is the use of synonyms (e.g. *Koning* and *Heer* may refer to the same card).

5. Experimental Setup

Our task is to determine a mapping from a spoken utterance to a semantic representation given in the form of semantic frame, where the frame structure is known in advance for the domain under consideration. The utterances are decoded into a sequence of AUDs, which are then mapped to semantic frames.

We present the experimental setups for the PATCOR dataset in detail, but employ an analogue approach to the DOMOTICA 3 dataset. The semantic frame structure for the PATCOR dataset is shown in Table 1. There are two possible frames: DealCard that takes no slots or slot values, and MoveCard that consists of six possible slots, each of them being composed of slot values. Not all the slots need to be inferred for each command. An example mapping from the spoken command to the corresponding semantic frame is given in Fig. 1. First we determine an acoustic representation of the spoken utterance in terms of the AUD sequence. Furthermore, we define a rule that maps the AUD sequence to a semantic frame, in this example $MoveCard(< FromSuit >, < FromValue >, < TargetSuit >, < TargetValue >)$. We present several different setups for our experiments.

5.1. Setup 1

For the setup 1 we define a set of first-order clauses for the mapping of AUD sequences to semantic frames:

$$HasAUD(+a, u) => FromSuit(+s, u)$$
$$HasAUD(+a, u) => FromValue(+v, u)$$
$$HasAUD(+a, u) => TargetSuit(+s, u)$$

Table 1: *PATCOR semantic frame structure*

Frame	Slot	Slot value
MoveCard	FromSuit	spades,diamonds,hearts,clubs
	FromValue	1,2,3,...,13
	TargetSuit	spades,diamonds,hearts,clubs
	TargetValue	1,2,3,...,13
	FromFoundation	1,2,3,4
	TargetFoundation	1,2,3,4
DealCard	{ }	{ }

$$HasAUD(+a,u) => TargetValue(+v,u)$$
$$HasAUD(+a,u) => FromFoundation(+f,u)$$
$$HasAUD(+a,u) => TargetFoundation(+f,u)$$
$$HasAUD(+a,u) => DealCard(u)$$

where $s \in \{spade, diamond, heart, club\}$, $v \in \{1, ..., 13\}$ and $f \in \{1, 2, 3, 4\}$. The $HasAUD(a, u)$ predicate is an evidence predicate, which states that a particular AUD *a* is part of the AUD sequence (utterance) *u*. Since we want to account for the co-occurrence of the successive AUDs, we let *a* be AUD bigrams as well. The predicates at the right side of the implication operator define the mapping of the AUD sequence *u* to a particular slot value. The + operator is a per constant operator that produces a separate clause for each combination of *a* (AUD unigram or bigram) and slot value. A separate weight is also learned for each clause obtained in this way. A part of the grounded MLN for one AUD belonging to the utterance *u* is presented in Fig. 2. Each node in this graph is a ground predicate obtained by assigning constants (e.g. $spades, diamonds, hearts, clubs$) to variables (e.g. s) in the FOL formulas above. An arc connects each two ground predicates that appear together in one grounding of a formula.

Finally, we infer the probabilities of mapping the AUD sequence to each of the slot values given in Table 1. The slot value with the highest probability is chosen for every slot, however only if it is higher than a predefined threshold; hence not all the slots need to be inferred for a semantic frame. For the example given in Fig. 1 slots $FromFoundation$ and $TargetFoundation$ stay below the threshold and hence these are not inferred.

5.2. Setup 2

We use the same set of FOL clauses as in setup 1 to define mappings, however, we introduce one additional *null* slot value *N* for each slot in Table 1, which is assigned to an utterance in the training dataset when there exists no mapping to a particular slot.

A major difference compared to setup 1 is the inference part; there is no need to use the threshold anymore, since all the slots are inferred for each semantic frame. However, we know that slots which are mapped to *null* slot value *N* shouldn't be inferred in the first place, so they are dropped before the evaluation.

5.3. Setup 3

In this setup mappings are learned in a hierarchical way, using a two-step approach. The first step defines the mappings of the AUD sequences to slots, which can be realized using a MLN

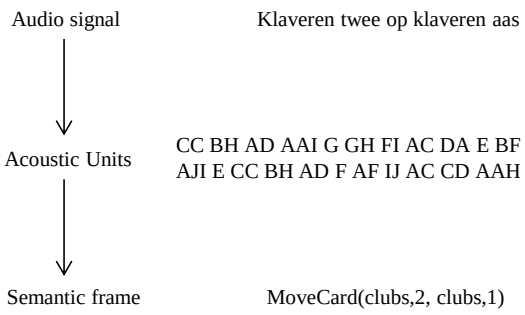

Figure 1: *Example mapping from the input speech to the semantic frame via an AUD sequence for PATCOR dataset*

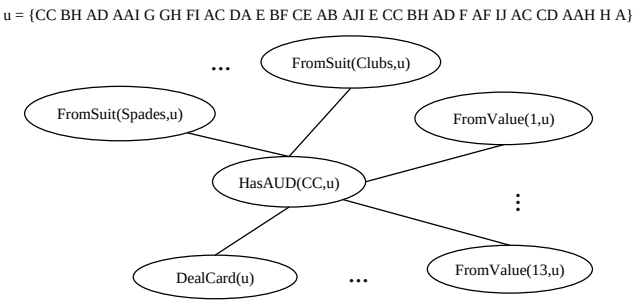

Figure 2: *A part of the grounded MLN for one AUD (CC) belonging to the utterance u*

with a single clause:

$$HasAUD(+a,u) => Command(+s,u)$$

where $s \in \{FromSuit, FromValue, TargetSuit, TargetValue, FromFoundation, TargetFoundation, DealCard\}$. After the inference is completed, the AUD sequences are divided into classes; one class per each inferred slot.

In the second step a separate MLN is constructed for each slot, which defines mappings of AUD sequences (extracted in the first step) to slot values that belong to the particular slot. As the inference is done separately, the obtained mappings for all the slots are joined before the evaluation.

6. Results and Discussion

For the evaluation we use the five-fold cross-validation procedure as in [27], where the dataset is partitioned into five subsets, four of them being used for training and the remaining one for testing. The cross-validation procedure is repeated 5 times (folds), with each of the subsets used exactly once as the test dataset. The performance measure is then averaged over all folds. As a performance measure we use the slot F-score, which is the harmonic mean of slot precision and slot recall. Slot precision is the number of correctly filled slots divided by the total number of filled slots in the induced semantic frame, while slot recall is defined as a number of correctly filled slots divided by the total number of filled slots in the reference semantic frame.

We use three baselines to compare the proposed MLN semantic analysis approach with. The first one is based on NMF, where the input speech is decomposed into recurrent acoustic patterns represented as AUDs, that are linked to semantic repre-

Table 2: *F-scores for different setups (PATCOR dataset)*

Speaker	1	2	3	4	5	6	7	8	Average
# Utterances	274	169	260	278	221	247	223	240	239
Baseline: NMF	66.1	69.3	76.2	55.9	90.9	54.7	77	48.5	67.3
Baseline: MNB	61.8	81.6	74.7	62	86.5	56.7	72.7	48.3	68
Baseline: SVM	59.7	80	77.8	57.5	89.4	49.9	65.9	45.1	65.7
MLN: Setup 1	66.3	82.9	80.6	65.1	91.4	54.8	79.1	56.5	72.1
MLN: Setup 2	68.4	85.6	83.6	67.4	94.5	65.1	81.4	56.1	75.3
MLN: Setup 3	68.4	83	83.2	67.6	93.5	66.1	82.2	56.2	75
MLN: Transcriptions	77.6	91.6	94.6	83.2	97.6	78	96.5	66.6	85.7

Table 3: *F-scores for different setups (DOMOTICA 3 dataset)*

Speaker	17	28	29	30	31	34	35	41	44	Average
# Utterances	347	204	174	198	225	331	268	144	164	228
Baseline: NMF	96.3	82	94.5	87.6	74.3	90.2	94.6	86.8	93.3	88.8
Baseline: MNB	94.1	79.3	88.8	85.2	73.8	91.4	95.3	86.2	96	87.8
Baseline: SVM	97.6	76.6	93.5	86.1	73.3	93	97.4	85.2	95.8	88.7
MLN: Setup 1	98.5	90.6	97	92.1	83.1	96.2	98.8	91.6	98.8	94.1
MLN: Setup 3	98.2	90	96.6	90	81.5	96.7	98.8	91.8	99	93.6
MLN: Transcriptions	100	100	100	100	100	100	100	100	100	100

sentations. This process is weakly supervised by labeling each utterance's acoustic representation in the training dataset with slot values to which the utterance refers to [27]. Additionaly, we use a classification approach based on a Multinomial Naive Bayes (MNB) model [28] and one based on Support Vector Machines (SVM) [29, 30]. Counts of AUDs unigrams and bigrams are used as features in MNB and SVM, where the feature value equals zero if the AUD does not appear in the utterance and equals the unigram or bigram count otherwise. The dimensionality of the feature vector is then reduced and only the 50% most relevant features for classification are kept, as measured by a mutual information criterion.

Results for the three different setups in terms of F-scores are given in Table 2 and Table 3 for the PATCOR and DOMOTICA 3 datasets, respectively. MLN-based semantic analysis utilizing setup 1 outperforms the baselines for almost all the individual speakers. On average, we get an absolute improvement in F-score of over 5% for the DOMOTICA 3 and over 4% for the PATCOR dataset.

Analyzing the inferred semantic frames we noted that a significant source of error for the setup 1 was the inference of unwanted additional slots (e.g. $FromFoundation$ and $TargetFoundation$ slots in the example shown in Table 1). It turned out to be very hard to define a reasonable rejection threshold, which would be a good trade-off between inferred and rejected slots within a frame. Hence, we employed a different approach (setup 2), where the hard threshold is avoided by mapping the spoken utterance to a null slot value for all the unwanted slots during the training phase. In this way we not only learn to which slots the utterance is mapped, we also learn to which ones it should not be mapped. This resulted in an additional absolute improvement in F-score of 3.2% on average for the PATCOR dataset. Setup 2 was not applied to DOMOTICA 3 since all the slots are always inferred for each frame, hence no improvement is possible in this way.

We also experimented with a hierarchical learning approach (setup 3) where we first learn mappings to slots, then subsequently for each slot learn mappings to its slot values. Although the F-scores slightly decrease for both datasets, the major benefit is the fact that the large task can in this way be divided into smaller subtasks. This relaxes one of the main drawbacks of the MLNs, i.e., the fact that for large tasks the inference is potentially very slow [15]. There is also a benefit in terms of computational complexity in the learning phase: learning time is decreased by 27%.

Finally, we give F-scores obtained on clean text transcriptions of the spoken utterances, which serve as the upper bound of what can be achieved on noisy input. Note that for the DOMOTICA 3 dataset we obtain an F-score of 100% since the ground truth mappings are unambiguous. On the other hand, the performance for the PATCOR dataset is quite a bit lower due to inconsistencies and inexactness of the speakers.

7. Conclusions

The results presented in this paper show that MLNs are a promising tool for the task of semantic analysis of spoken language, even in the presence of noisy and inconsistent input data. Coupled with an unsupervised learning of speech representations the approach is especially applicable to smaller and noisy domains, such as home automation task for speech-impaired users where standard ASR techniques do not perform well. The experimental results show a clear improvement over three baseline systems. However, a comparision with the F-score achieved on error-free text transcriptions indicates that there is still significant room for further improvement.

8. Acknowledgements

This work was partly funded by DFG, contract no. Ha 3455/9-1, within the Priority Program SPP1527 Autonomous Learning. V. Despotovic was supported by an Erasmus Mundus Action 2 scholarship within the EUROWEB scholarship programme. We wish to thank J. Gemmeke and H. Van hamme from KU Leuven for making available to us the datasets used in this paper.

9. References

[1] B. Coppola, A. Moschitti, and G. Riccardi, "Shallow semantic parsing for spoken language understanding," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, ser. NAACL-Short '09, 2009, pp. 85–88.

[2] Y.-Y. Wang, "Strategies for statistical spoken language understanding with small amount of data - an empirical study." in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan*, 2010, pp. 2498–2501.

[3] R. D. Mori, F. Béchet, D. Hakkani-Tür, M. McTear, G. Riccardi, and G. Tur, "Spoken language understanding—interpreting the signs given by a speech signal," *IEEE Signal Processing Magazine*, pp. 50–58, 2008.

[4] E. Sanders, M. B. Ruiter, L. Beijer, and H. Strik, "Automatic recognition of dutch dysarthric speech: a pilot study," in *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA*, 2002.

[5] J. F. Gemmeke, J. V. D. Loo, G. D. Pauw, J. Driesen, H. V. hamme, and W. Daelemans, "A self-learning assistive vocal interface based on vocabulary learning and grammar induction," in *Proc. INTERSPEECH*, 2012, pp. 1–4.

[6] J. F. Gemmeke, B. Ons, H. Van hamme, J. van de Loo, W. D. G. De Pauw, J. Huyghe, J. Derboven, L. Vugen, B. van Den Broeck, P. Karsmakers, and B. Vanrumste, "Self-taught assistive vocal interfaces : An overview of the ALADIN project," in *Proc. INTERSPEECH*, 2013, pp. 1–5.

[7] O. Walter, V. Despotovic, R. Haeb-Umbach, J. Gemmeke, B. Ons, and H. Van hamme, "An evaluation of unsupervised acoustic model training for a dysarthric speech interface," in *INTERSPEECH 2014*, 2014. [Online]. Available: http://nt.uni-paderborn.de/public/pubs/2014/WaDeHaebGeOnVa14.pdf

[8] Z. Chen, S. Tamang, A. Lee, X. Li, M. Passantino, and H. Ji, "Top-down and bottom-up: A combined approach to slot filling." in *AAIRS*, ser. Lecture Notes in Computer Science, vol. 6458, 2010, pp. 300–309.

[9] S. Riedel and I. Meza-Ruiz, "Collective semantic role labelling with markov logic," in *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, ser. CoNLL '08, 2008, pp. 193–197.

[10] I. Meza-Ruiz and S. Riedel, "Multilingual semantic role labelling with markov logic," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, June 2009, pp. 85–90.

[11] W. Che and T. Liu, "Jointly modeling wsd and srl with markov logic," in *Proceedings of the 23rd International Conference on Computational Linguistics*, ser. COLING '10, 2010, pp. 161–169.

[12] M.-J. Meurs, F. Duvert, F. Lefevre, and R. D. Mori, "Markov logic networks for spoken language interpretation," *Information Systems Journal*, pp. 535–544, 2008.

[13] C. Kennington and D. Schlangen, "Markov logic networks for situated incremental natural language understanding." in *SIGDIAL Conference*, 2012, pp. 314–323.

[14] H. Poon and P. Domingos, "Unsupervised semantic parsing," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, ser. EMNLP '09, 2009, pp. 1–10.

[15] C. R. Kennington and D. Schlangen, "Situated incremental natural language understanding using markov logic networks," *Computer Speech and Language*, vol. 28, no. 1, pp. 240–255, 2014.

[16] T. Netsanet, O. Bart, van de Loo Janneke, G. Jort, D. P. Guy, D. Walter, and V. hamme Hugo, "Metadata for corpora patcor and domotica-2," KU Leuven, Tech. Rep., 2013.

[17] M. Richardson and P. Domingos, "Markov logic networks," *Mach. Learn.*, vol. 62, no. 1-2, pp. 107–136, 2006.

[18] S. Kok, M. Sumner, M. Richardson, P. Singla, H. Poon, D. Lowd, J. Wang, and P. Domingos, "The alchemy system for statistical relational AI," Department of Computer Science and Engineering, University of Washington, Seattle, WA., Tech. Rep., 2009.

[19] H. Poon and P. Domingos, "Sound and efficient inference with probabilistic and deterministic dependencies," in *Proc. of the 21st National Conference on Artificial Intelligence (AAAI '06), Boston, Massachusetts, USA*, 2006, pp. 458–463.

[20] S. Chaudhuri and B. Raj, "Unsupervised structure discovery for semantic analysis of audio," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, Nevada, United States.*, 2012, pp. 1187–1195.

[21] S. Chaudhuri, M. Harvilla, and B. Raj, "Unsupervised learning of acoustic unit descriptors for audio content representation and classification." in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy*, 2011, pp. 2265–2268.

[22] M. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervised Training of an HMM-Based Self-Organising Unit Recognizer with Applications to Topic Classification and Keyword Discovery," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 210–223, Jan. 2013.

[23] O. Walter, V. Despotovic, R. Haeb-Umbach, J. Gemmeke, B. Ons, and H. Van hamme, "An evaluation of unsupervised acoustic model training for a dysarthric speech interface," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore*, 2014, pp. 1013–1017.

[24] O. Walter, T. Korthals, R. Haeb-Umbach, and B. Raj, "Hierarchical System for Word Discovery Exploiting DTW-Based Initialization," in *Automatic Speech Recognition and Understanding Workshop (ASRU 2013)*, Dec. 2013, pp. 386–391.

[25] J. F. Gemmeke, B. Ons, N. Tessema, H. V. hamme, J. van de Loo, G. D. Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. V. D. Broeck, P. Karsmakers, and B. Vanrumste, "Self-taught assistive vocal interfaces: an overview of the ALADIN project," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013, pp. 2039–2043.

[26] C. Middag, "Automatic analysis of pathological speech," Ph.D. dissertation, Ghent University, Belgium, 2012.

[27] B. Ons, N. Tessema, J. van de Loo, J. Gemmeke, G. D. Pauw, W. Daelemans, and H. V. hamme, "A self learning vocal interface for speech-impaired users," in *4th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Lyon, France*, 2013, pp. 78–81.

[28] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in *Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence, Cairns, Australia*. Berlin, Heidelberg: Springer-Verlag, 2004, pp. 488–499.

[29] C. Cortes and V. Vapnik, "Support-vector network," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[30] T. Joachims, "Text categorization with suport vector machines: Learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning*, ser. ECML '98. London, UK, UK: Springer-Verlag, 1998, pp. 137–142.