Acoustic Correlates of Meaning Structure in Conversational Speech

A. V. Ivanov[1], *G. Riccardi*[1], *S. Ghosh*[1], *S. Tonelli*[2], *E. A. Stepanov*[1]

[1]DISI, University of Trento, Italy
[2]FBK-IRST, Trento, Italy
{ivanov, riccardi, ghosh, stepanov}@disi.unitn.it[1], satonelli@fbk.eu[2]

Abstract

We are interested in the problem of extracting meaning structures from spoken utterances in human communication. In Spoken Language Understanding (SLU) systems, parsing of meaning structures is carried over the word hypotheses generated by the Automatic Speech Recognizer (ASR). This approach suffers from high word error rates and ad-hoc conceptual representations. In contrast, in this paper we aim at discovering meaning components from direct measurements of acoustic and non-verbal linguistic features. The meaning structures are taken from the frame semantics model proposed in FrameNet, a consistent and extendable semantic structure resource covering a large set of domains. We give a quantitative analysis of meaning structures in terms of speech features across human–human dialogs from the manually annotated LUNA corpus. We show that the acoustic correlations between pitch, formant trajectories, intensity and harmonicity and meaning features are statistically significant over the whole corpus as well as relevant in classifying the target words evoked by a semantic frame.

Index Terms: spoken language understanding, spoken dialog, frame semantics, speech mining, acoustic features

1. Introduction

We are interested in the problem of extracting meaning structures from spoken utterances in human communication. In Spoken Language Understanding (SLU) systems, parsing of meaning structures is carried over the word hypotheses generated by the Automatic Speech Recognizer (ASR)[1]. The automatic transcripts generated by the ASR are parsed and syntactic/semantic chunks are extracted. Such parsing models are either hand-crafted (e.g. semantic grammars) or statistically trained from annotated corpora with ad-hoc and application specific concept labels. This computational model has had success in applications such as spoken dialog systems but may be limited by semantic coverage or high word error rates, in the case of unconstrained conversational systems. In this paper we aim at discovering meaning components from direct measurements of acoustic and non-verbal linguistic features. Such components include the *most* semantically important word as well as its dependents within the semantic structures associated to a spoken utterance.

This approach to speech understanding is motivated by relevant research in speech and language processing, phonetics and language acquisition. In language acquisition, the most important questions are how to acquire words, their meaning while interacting in a physical and social context. In [2, 3] , meaning is grounded into machine actions and no semantic structure *bias* is assumed or exploited. In [4], meaning is directly learned from phone sequence distributions and visual features in the context of infant-directed speech. In computational linguistics the role of prosodic features to predict phrase structures has been well studied for cue phrases [5] and classification of intonational phrase boundaries [6]. Prosodic patterns have been also used as features for detecting and classifying dialog acts in conversational speech [7]. More recently the use of prosodic information has been applied to speech summarization [8]. From an acoustic point of view, prosody has been shown to manifest in variation of pitch, loudness, segment durations and specific manner of articulation.

Theories have been proposed to explain the way prosody can convey meaning through variation of these parameters. In [9, 10] three *codes* characterize prosodic patterns. The *frequency* (or size), *effort* and *production* code. Following [11] there is a link between paralinguistic assertion and lowering ones pitch. This link has profound origins as larger species, which are often perceived to have dominant or dangerous behavioral pattern, normally have larger vocal apparatus and produce lower pitch.

In section 2 we describe the meaning structures we aim at grounding into acoustic and linguistic features of the spoken utterances. Here we also describe how such model has been used to annotate the human-human dialog corpus, with a summary statistics of data split, used further in classification experiments. In section 3 we thoroughly exploit the acoustic features in order to use those in acoustic prediction. The target word classification with lexical features is described in section 4. Then a combined classification measurements with oracle accuracy is presented in section 5. We finally conclude in section 6.

2. The FrameNet Semantic Structures

We carried out our experiments using the LUNA spoken dialog corpus, which was developed in the context of the LUNA research project for next-generation spoken dialog interfaces ([1]) and was manually annotated with a multi-layered approach, including attribute-value information, Predicate-Argument-Structure (PAS) and dialog acts ([12]). This corpus includes human–human (HH) dyadic conversations of Italian speakers engaged in a problem-solving task in the domain of software/hardware troubleshooting, whereas the human–machine (HM) dialogs were acquired with a Wizard of Oz approach (WOZ) for the problem specification task only. In the corpus preparation phase, we extracted for each token the lemma, the turn id, the time-stamp and also the PAS label when available. PAS annotation was carried out applying the FrameNet paradigm as described in [13]. This annotation model covers a set of prototypical situations called *frames*, the frame-evoking words called *lexical units* or *target words* and the roles or participants involved in these situations, called *frame elements* (FEs). The latter are typically the syntactic dependents of the lexical units. All lexical units belonging to the same frame have similar semantics and valence (for details about the annotation scheme, see [12]). We adopted where possible the frame

and frame element labels which were originally defined for the English FrameNet project. Some new frame definitions were introduced only in case of missing elements in the off-the-shelf resource.

An example annotation of two dialog turns is reported in Fig. 1. For each Italian utterance transcription, we annotate the target words, the frame and its frame elements. The target words (in bold) *dire* and *chiamo* are assigned to a frame label in capitals, resp. TELLING and BEING_NAMED. This means that *dire* evokes the prototypical situation that is defined as TELLING in the FrameNet database, while *chiamo* evokes a situation called BEING_NAMED. Given that a frame is also characterized by some frame elements or semantic roles, *Addressee* and *Message* are the FEs expressed in the first utterance for TELLING, and *Entity* and *Name* in the second one for BEING_NAMED.

Figure 1: *PAS Annotation: an example.*

In this work, we focus primarily on the annotation of *target words*, and in particular on the criteria for identifying target words in a turn. In the early stages of the Berkeley FrameNet project[1], one frame per sentence was annotated, so just one target word was chosen in every sentence. More recently, the Berkeley group has started also another annotation effort, called *continuous-text annotation*, in which all possible valence-bearing words in a sentence are annotated as target words. In LUNA, we adopted an intermediate approach, following the idea that all *semantically relevant* target words with a syntactic subcategorization pattern have to be identified and annotated, possibly skipping the utterances with empty or fragmentary semantics (e.g. disfluencies). As expected, most of the targets annotated with our approach are verbs (almost 71% of the occurrences, while 14% are nouns and the rest adjectives and adverbs). In the FrameNet database, instead, the occurrences of verbal targets w.r.t. other PoS are more evenly distributed (44% verbs, 39% nouns, 16% adjectives), since the annotated sentences were selected in order to be representative of different *frames*, thus they are more balanced.

In order to assess the relation of the different annotation levels in the LUNA corpus, we performed the alignment of the multiple layers, viz. annotation of tokens, turns and PAS for 125 HH dialogs, mapping each token with turn ID and timestamp as well as with target / non-target) label. A summary of the corpus statistics is reported in Table 1.

3. Acoustic features

Audio recordings of the LUNA spoken dialog corpus were recorded as a mixed duplex channel with 8 KHz mono 16-bit pulse-coded modulation (PCM). The recordings were segmented into speech dialog turns and transcribed by human annotators. Afterwards, these turns were annotated, as described in the previous section. For the purposes of the experiment, all words in the recordings were labeled either as "Non-target word", "Target word" or "Frame element".

[1] http://framenet.icsi.berkeley.edu/

Table 1: *LUNA corpus statistics for the training, development and test sets.*

	Train	Devel	Test
No. of dialogs used	94	11	20
No. of utterances used	4748	506	1131
Average no. of utterances per dialog	50.51	46.00	56.55
Total no. of tokens	34123	3479	7912
Average utterance length (in tokens)	7.19	6.88	7.00
Average dialog length (in mins)	3.21	3.23	3.39
No. of unique tokens	3307	872	1388
No. of lemmas	2312	688	1017
No. of PoS tags	24	17	15
No. of unique frames	204	107	135

We have passed each individual turn through the forced alignment procedure with the Italian language ASR trained with the Sphinx-3 toolkit on the LUNA corpus.

We have extracted measurements of speech pitch (F_0), formant trajectories ($F_x, x = 1, 2, 3$), intensity (I_{tot}) and harmonicity (I_{hnr}) with the standard algorithms provided in the PRAAT toolkit [14]. We have performed all measurements over a signal window of 40 ms with the frame rate of 100 Hz. The latter two measurements were combined to obtain an estimation of the intensity of a harmonic component (I_{harm}) of the speech signal. Employment of I_{harm} was motivated by the possible effect of acoustic interferences like environmental noise or other non speech-like sounds. Intensity of a harmonic component is also believed to better correspond to the intensity of phonation and paralinguistic stress. It is less distorted with wide-band energy bursts of plosives or fricatives. We have used the following formula for the depicted combination:

$$I_{harm} = I_{tot} + I_{hnr} - 10\log_{10}(10^{I_{hnr}/10} + 1). \quad (1)$$

Note that when I_{hnr} is high then $I_{harm} \approx I_{tot}$. However, when I_{hnr} is sufficiently negative, then I_{harm} can also become negative.

The segmentation resulting from the forced alignment was then used to extract token-specific estimates. These absolute values were then compared to an average value of the given measurement throughout a whole turn. The measurements performed that way allow for a direct verification of the "code"-theories. A statistical analysis of the relative features has revealed that there exists a statistically significant difference between the mean values of the segmental relative features depending on which role a given segment possesses.

The average deviation of the maximal intensity of the harmonic component attained within a given segment from the average speech harmonic component intensity of the whole turn is higher for target words as opposed to frame elements and non-target words (see Fig. 2 for further detail). This observation is predicted by the effort code. It is notable that a maximal–to–mean intensity margin of the harmonic component has a clearance of 8 db between the target and all other words. An identically measured margin for an entire signal intensity I_{tot} does not exceed 1.5 db. This fact confirms our conjecture that the harmonic component intensity represents the paralinguistic stress pattern in a better way.

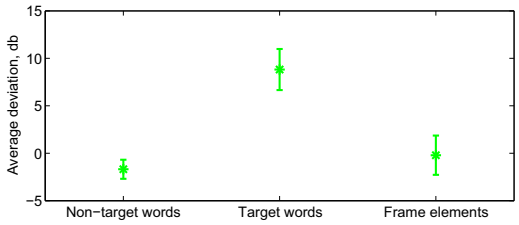

Figure 2: Average deviation of the maximal within-segment harmonic intensity from the corresponding average of the whole turn (Y-axis).The measurements performed for all label types (X-axis). Confidence intervals are given for p-value $p = 0.05$.

The average deviation of the minimal pitch frequency of the voiced interval attained within a given segment from the average pitch frequency of the whole turn is lower for target words as opposed to others (see Fig. 3 for further detail). The observation is in agreement with the frequency code.

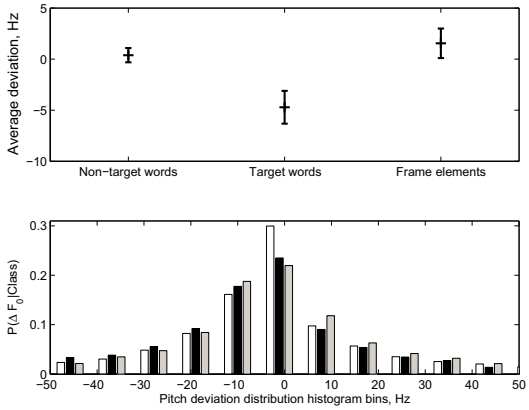

Figure 3: Upper Plot: Average deviation of the minimal within-segment pitch frequency from the corresponding average of the whole turn (Y-axis). The measurements performed for all label types (X-axis). Confidence intervals are given for p-value $p = 0.05$. Lower Plot: Corresponding distributions $P(\Delta F_0|Class)$, black – target words, white – non-target words, gray – frame elements.

The pitch dynamic range of target words is approximately 15 Hz larger in comparison to the average dynamic range of all of the words in that turn. This observation is in agreement with the effort code. The measurement is statistically significant with $p = 0.05$.

The average deviation of the mean duration of the voiced interval within a given segment from the average duration of the voiced intervals in the whole turn is larger for target words as opposed to frame elements and non-target words. At this point we did not reach a conclusion if voicing duration represents an independent feature or is a byproduct of the generally increased intensity of the harmonic speech component. The PRAAT performs pitch measurements through the use of a correlation statistics. Allegedly, it is able to uncover and track more intense harmonic structures from larger distances. We have additionally performed an analysis of the duration of the individual phonemes as it was recorded during the forced alignment. But we have not found a consistent pattern, that depends on the role of the segment under consideration.

The average inter-frame formant frequency difference (F_2 and F_3) is larger in comparison to the utterance mean for target words as opposed to frame elements and non-target words (see Fig. 4 for further detail). The formant dynamics is in agreement with the effort code. However, we expect the formant features to be more informative if this formant dynamics gets conditioned on the particular phoneme that is being uttered.

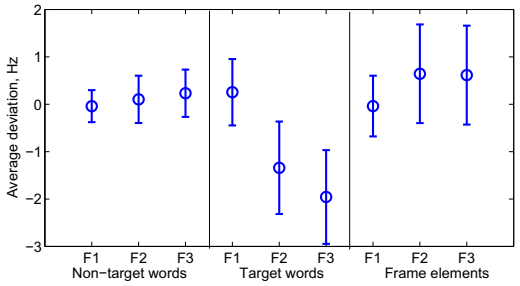

Figure 4: Average deviation of the average interframe formant frequency deltas from the corresponding utterance average (Y-axis). The measurements performed for three first formants "F1", "F2", "F3". Confidence intervals are given for p-value $p = 0.05$.

4. Target Word Classification with lexical features

The classification experiment involved identifying each token of a presegmented utterance as either a target word or a non-target word i.e. a binary classification task. This task was done using lexico-syntactic features: *(a) Part-of-Speech (PoS) tags* automatically added with the Chaos parser [15]. *(b) Lemma information*, annotated with TreeTagger[16]. Human PAS annotation is primarily dependent on these two types of information for target / non-target word classification. Two further features comprise *(c) lowercased token* and its *(d) previous token*, including utterance boundary information.

The experiment was carried out using BoosTexter[17] classifier, that uses AdaBoost algorithm, which is initialized with a set of weak hypotheses by calling these weak classifiers in a series of iterations, and finally combining the weak hypotheses into a single rule.

For this task, baseline was established by using only tokens as feature; the number of iterations was optimized for minimum false rejection rate, determined by best performance over the development data. To evaluate the classification performance, test output was scored using precision, recall, and $F1$-measure. Table 2 shows all results produced as baseline, single features and features in combination.

Table 2: *Results with Baseline, Single and Combined Features*

Features	Precision	Recall	F1-measure
Baseline (token)	0.759	0.648	0.699
PoS	0.655	0.825	0.730
Lemma	0.764	0.747	0.755
Token+PoS	0.787	0.800	0.793
Token+lemma+PoS	0.797	0.803	0.800
Token+Prev_tok+PoS	0.765	0.857	0.808
Token+Prev_tok+lemma+PoS	0.782	0.841	0.810

We observe that combined features: lowercased tokens, lemmas and PoS tags already achieve a better performance compared to baseline, since all these in combination convey a good amount of information about target words. The "previous token" feature adds a context to the combined classifier (i.e. the lowercased token, its lemma and PoS tag). Thus, the result is further improved using all four features in combination.

5. Combination of Lexical and Acoustic features

The effectiveness of the acoustic measurements in predicting target word classification task has been evaluated in combination with the lexical features. A multilayer perceptron (MLP) and a support vector machine (SVM) were used as classifiers for the acoustic features driven classifier. The MLP had one hidden layer with only 200 neurons. The output layer had two neurons corresponding to the two class labels that are being trained in the supervised way - "a target word" and "not a target word". The SVM was using a linear kernel. The feature vector contained all of the features depicted above.

The resulting classifiers on the test set were able to attain a performance level being $F1 \approx 0.3747$ for MLP and $F1 \approx 0.3397$ for SVM. The chance performance on the same test set has $F1 \approx 0.2972$. Thus, it is possible in principle to use acoustic information to infer semantics of a given segment. As is illustrated by Fig. 3, the histograms of distributions of features are almost overlapping. It is a large number of experiments that allows us to record a statistically significant shift in the mean values of the features. We expect that better employment of a turn context may improve recognition performance.

Another possibility is to integrate results of acoustic classification over multiple instances of the same word. This leads to a potential application of the acoustic classifier in automated language acquisition. The words, and in general the whole linguistic contexts, which are consistently being marked as the frame-generating targets by the acoustic classifier may be incorporated into a model of the linguistic classifier, thus enabling an autonomous acquisition of the linguistic model from the spoken data only.

Table 3: *Performances of the best lexical and acoustic feature based classifiers and their oracle performances on the target word classification task .*

Classifier	Prec.	Recall	F1
Lexical Features	0.782	0.841	0.810
Acoustic Features	0.247	0.774	0.375
Oracle Combination	0.935	0.913	0.924
Baseline Linguistic Classifier	0.759	0.648	0.699
Oracle Comb. (+ best acoustic)	0.926	0.811	0.865

As shown in Table 3, the figure of merit of combined systems could be as high as $92,4\%$. The combination of both acoustic and linguistic classifiers has the potential to be very accurate.

6. Conclusion

In the experiments with large amounts of spoken data we have observed a statistically significant deviation of the means of objectively measured segment parameters depending on the meaning of that segment. As such, our observation confirms our initial conjecture regarding the acoustic features grounding of the semantic elements within an utterance. Our findings support the theories of speech prosody and the effects of frequency and effort codes. An other important result is the correlation between acoustic measurements and a semantic representation of meaning that is linguistically motivated and consistent across domains. The preliminary classification experiments of the fine semantic structure elements are very encouraging and motivate the combination of acoustic and lexical features.

7. Acknowledgements

This work was partially supported by the European Commission Marie Curie Excellence Grant for the ADAMACH project (contract No. 022593).

8. References

[1] R. De Mori, F. Béchet, D. Hakkani-Tür, M. McTear, G. Riccardi, and G. Tür, "Spoken Language Understanding: A Survey," *IEEE Signal Processing*, vol. 25, no. 3, 2008.

[2] A. L. Gorin, S. E. Levinson, and A. Sankar, "An experiment in spoken language acquisition," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 224–240, January 1994.

[3] A. L. Gorin, D. Petrovska-Delacretaz, G. Riccardi, and J.H. Wright, "Learning spoken language without transcription," in *Proceedings of IEEE ASRU Workshop*, 1999.

[4] D.K. Roy and A. P. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science: A Multidisciplinary Journal*, vol. 26, pp. 113–146, 2002.

[5] J. Hirschberg and D. Litman, "Empirical studies on the disambiguation of cue phrases," *Computational Linguistics*, vol. 19, pp. 501–530, Spetemeber 1993.

[6] M.Q. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Computer Speech and Language*, vol. 6, pp. 175–196, 1992.

[7] A. Stolcke and et al., "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, pp. 339–373, Spetemeber 2000.

[8] S. Maskey and J. Hirschberg, "Summarizing Speech Without Text Using Hidden Markov Models," in *Proc. of HLT-NAACL'2006*, June 2006.

[9] C. Gussenhoven, "Intonation and interpretation: phonetics and phonology," in *Proc. of Speech Prosody*, 2002, pp. 47–57.

[10] B. Post, D. Brechtje, and C. Gussenhoven, "Fine Phonetic Detail and Intonational Meaning," in *Proc. of 16th Int. Congress of Phonetic Sciences (ICPhS)*, August 2007, pp. 191–196.

[11] J.J. Ohala, "The frequency code underlines the sound symbolic use of voice of pitch," in *Sound symbolism*, L. Hinton et al., Ed., pp. 325–347. Cambridge: Cambridge Univ. Press, 1994.

[12] M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi, "Annotating Spoken Dialogs: from Speech Segments to Dialog Acts and Frame Semantics," in *Proc. of* 2^{nd} *Workshop on Semantic Represent. of Spoken Lang.*, 2009.

[13] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet Project," in *Proc. of ACL/Coling*, 1998.

[14] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[15] Roberto Basili and Fabio Massimo Zanzotto, "Parsing Engineering and Empirical Robustness," *Journal of Natural Language Engineering*, June 2002.

[16] G. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Proceedings of International Conference on New Methods in Language Processing*, 1994.

[17] Robert E. Schapire and Yoram Singer, "Boostexter: A boosting-based system for text categorization," in *Machine Learning*, 2000, pp. 135–168.