Proc. of Speech Prosody 2010, Chicago, May 2010,4 pages, no pagination

Word Accent and Emotion

Dino Seppi[1], *Anton Batliner*[2], *Stefan Steidl*[2], *Björn Schuller*[3], *Elmar Nöth*[2]

[1]ESAT, Katholieke Universiteit Leuven, Belgium
[2]Pattern Recognition Lab, University of Erlangen-Nuremberg, Germany
[3]Institute for Human-Machine Communication, Technische Universität München, Germany
dino.seppi@esat.kuleuven.be

Abstract

In this paper, we address the question whether prosodically/linguistically prominent syllables carrying the word accent (stressed syllables) are better indicators for emotional marking than unstressed syllables. To this aim, we use a large spontaneous database with children interacting with Sony's Aibo robot, annotated with word-based emotion labels, large acoustic-prosodic feature vectors, and support vector machines as classifiers. It turns out that, in most of the cases, stressed syllables are better emotion markers than unstressed syllables. Moreover, we discuss specific phenomena such as vocatives and other constellations, to be modelled in future studies.
Index Terms: emotion, linguistics, paralinguistics, word accent, lexical stress, automatic classification

1. Introduction

At first sight, the question that we will pursue in this paper seems to be a strange one: whether emotion[1] conveyed via speech is a linguistic phenomenon or not – it is common wisdom that emotion is part of paralinguistics. At second sight, it is evident that at least those emotions that are signalled with written language in texts or that can be detected in speech without any reference to acoustics, are 'linguistic phenomena': either they *manifest* themselves via language (in object-language) or at least they are *talked about* in meta-language. Note that here, we do not necessarily want or need to incorporate emotions as part of linguistics the same way as semantics or pragmatics are parts of linguistics; we rather aim at the means that are used by humans to signal these phenomena, and by that, at the features and units we are using when trying to detect emotions. Whereas the situation is evident when we deal with written language – there is no acoustics whatsoever – we can even deal with speech, trying to detect emotions, without using acoustic features for this very task itself: we simply can use automatic speech recognition (ASR) to recognize the spoken word chain, and only syntactic/lexical information (e. g. bag-of-words or n-grams) for recognizing emotions. Most likely, we will produce some word errors, but as far as we can see, emotion recognition does not necessarily deteriorate [1].

The impression that the signalling of emotion in speech is foremost done with (global) acoustics might be caused by the long prevailing experimental paradigm: researchers were using an identical sentence in the lab and let their subjects produce different emotions by changing acoustic parameters (prosody, tone of voice, etc.). In such a setting, and normally even when more realistic speech has been used, longer units such as sentences/turns/utterances/dialogue moves have been employed in emotion processing and were modelled as such.

The database used in the present study has been annotated on the word level, and several experiments have been conducted using words [2], turns, or chunks as units [3]. In a few other studies, alternative types of sub-turn (sub-word) units have been investigated, for instance voiced segments (sort of pseudo-syllable) in [4], and absolute time intervals, e. g. 500 ms., or relative time intervals (fixed number of segments per unit with equal length, e. g. three thirds in [5]), cf. as well [6]. Common to such alternative approaches is that they are linguistically 'blind' even if automatically detected voiced segments of course have much in common with syllables.

Accentuation is a typical linguistic/prosodic means to signal different words, or to tell apart different meanings of longer stretches of words: word accent (lexical stress) is – at least in languages such as English and German – a means to tell apart words such as 'SUBject vs. sub'JECT, and phrase accents signal different focal structures etc. Leaving aside contrastive accents, phrase accents are always manifested on word accent position. Most important might be the role of accentuation in attentional control and in structuring speech.

We will have a look at the signalling of emotion on word accent (+WA) syllables vs. unaccentuated (-WA) syllables. Our research interest can be formulated as null vs. alternative hypothesis. The null hypothesis (H0) claims that the signalling of emotion is simply modulated onto the speech flow, without telling apart +/-WA; thus classification performance should be roughly the same, no matter whether Words, or +WA, or -WA syllables are used. The alternative hypothesis (H1) claims that the signalling of emotion has much to do with linguistic structure and by that, it is more pronounced on +WA syllables than on -WA syllables. In Sec. 5 we discuss whether this alternative hypothesis can hold or whether there is a simpler explanation.

2. Material and annotation

The database used is a German corpus with recordings of children communicating with Sony's AIBO pet robot; it is described in more detail in [3] and other papers quoted therein. The children were led to believe that the AIBO was responding to their commands, whereas the robot was actually being controlled by a human operator who caused the AIBO to perform a fixed, predetermined sequence of actions; sometimes the AIBO behaved disobediently, thereby provoking emotional reactions. The data was collected at two different schools from 51 children (age 10 - 13, 21 male, 30 female; about 8.9 hours of speech without pauses, sampling rate 16 bits at 16 kHz).

The recordings were segmented automatically into 'turns'

[1]Note that emotion is used here in a broad sense, encompassing the 'proto-typical', 'big' emotions as well as other affective, emotion-related states such as stress, boredom, interest, etc.

using a pause threshold of 1000 ms. Five labellers listened to the turns in sequential order and annotated each word as neutral (default) or as belonging to one of ten other classes. If three or more labellers agreed, the label was attributed to the word (majority voting MV). The number of cases with MV is given in parentheses: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy* i. e. irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*, i. e. non-neutral but not belonging to the other categories (3), *neutral* (39169). 4707 words had no MV; all in all, there were 48401 words. As some of the labels are very sparse, they are mapped onto main classes [3]: *touchy* and *reprimanding*, together with *angry*, are mapped onto **A***ngry* as representing different but closely related kinds of negative attitude. (**A***ngry* can consist, for instance, of two *touchy* and one *reprimanding* label; thus the number of **A***ngry* cases is far higher than the sum of *touchy*, *reprimanding*, and *angry* MV cases.) Some other classes, like *joyful*, *surprised*, *helpless*, *bored*, and *rest* do not appear in this subset.

In this study, we restrict to a limited, emotionally rich subset of the whole database consisting of 4543 chunks (13202 words). Chunks are obtained by manually segmenting utterances using coarse syntactic and prosodic labels (cf. [3] for more details) and by selecting those chunks with at least one MV word. Subsequently, chunks are automatically segmented into words and syllables by Viterbi alignment [7]; the vocabulary consists of 525 words (510 syllables).

This selection resulted in a 4-class dataset consisting of 1772 words for **A***ngry*, 1238 words for **M***otherese*, 2452 words for **E***mphatic*, and 7740 for **N***eutral*. Note that we did not downsample frequent classes like **N***eutral* in an attempt of respecting as much as possible the skewness of the original distribution of the (remaining) labels. Also note that **E***mphatic* has to be conceived as a pre-stage of anger because on the valence dimension, it lies between neutral and anger [2, 3]; this is context-dependent and cannot be decided upon in any generic way, without knowing the data. However, **E***mphatic* is closest to **N***eutral* and thus somehow in-between a clear emotional state and a linguistic-prosodic phenomenon. Finally, syllable labels are assigned by directly assigning the respective word labels. Details on syllable and word statistics of this subset can be found in Tab. 1. In Fig. 1, the histogram of words per length (expressed in number of syllables) and per emotion is drawn. Especially for monosyllabic words, here we – somehow counterfactually – assume that they are all +WA words; we will come back to this topic in Sec. 5. Note that 176 words are truncated and the accented syllables are lost.

Table 1: *Syllable (*-WA*,* +WA*, and* +/-WA*) and* Word *statistics per emotion. 176 truncated words are missing the accented syllable.*

emotion	-WA	+WA	+/-WA	Word
M	370	1226	1596	1238
N	2867	7588	10455	7740
E	924	2450	3374	2452
A	1307	1762	3069	1772
	5468	13026	18494	13202

3. Experiments

In this paper, we will address three different classification problems. First, the 2-class problem **E***mphatic* vs. **N***eutral* (**EN**); by that, we can 'simulate' the more traditional linguistic/prosodic problem [± PROMINENCE]. This linguistic phenomenon is of course not the same as the type of emphasis found here, serving as a sort of 'pre-stage' to negative emotion in emotional speech; however, both phenomena have quite a lot in common although we are not aware of any study that systematically has investigated their relationship. Content words, e. g., are more prone to be linguistically prominent and at the same time to be marked 'emotionally' than function words.

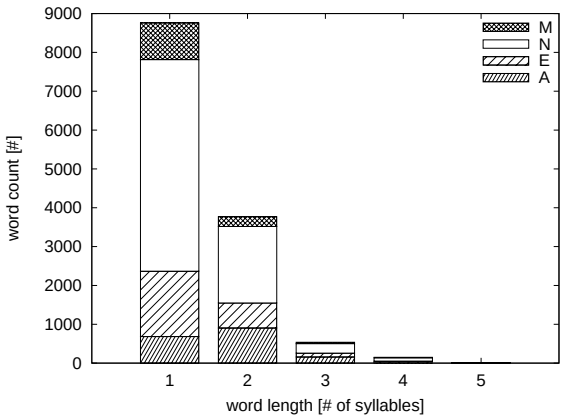

Figure 1: *Histogram of words per length [# of syllables], per emotion.*

Second, we address the 3-class problem **A***ngry*, **N***eutral*, and **M***otherese* (**AMN**); by that, we model a typical 'emotion constellation', i. e. valence with a negative, a neutral, and a positive state.

Third, we address the full 4-class problem described above (**AMEN**), i. e. we detail negative valence by telling apart the pre-stage **E***mphatic* and the main class **A***ngry*; this can be called a realistic modelling because it is based on all data, i. e. on all phenomena found to be relevant for emotion processing.

In the following we tackle each one of these classification problems and analyze the result of using different word subunits: as mentioned already, the speech signal has been segmented by forced alignment into words and syllables. In the experiments we consider words and syllables, which are either -WA, or +WA, or both -WA and +WA (i. e.+/-WA). As we are interested in the discrimination power of different (sub-) word units, other important factors should be normalized, such as the number of items. If it is true that the number of +WA is equal to the number of Word, for our data -WA is far less numerous, while +/-WA is (almost, cf. Sec. 2) the union of +WA and -WA patterns (cf. Tab. 1). To avoid the influence of the heterogeneous number of items in the classification experiments, we always randomly down-sampled the number of training patterns to the multiplicity of the -WA subset. Therefore, the results of Fig. 2 are obtained by training on (almost) the same amount of data per class. Another important variable to consider is the linguistic content: given the quite small (syllable and word) lexicon (cf. Sec. 2), chances are that the classifier is actually learning the linguistic content in the acoustic patterns. For this reason, we broke the feature set into two parts: the first group encompasses prosody-based features alone, while the second one encloses spectral features only (MFCC). We expect that the prosodic features are more independent of the linguistic content.

As just anticipated, a purely acoustic feature set is adopted for this study. We chose the most common and at the same time promising feature types and functionals covering prosodic and

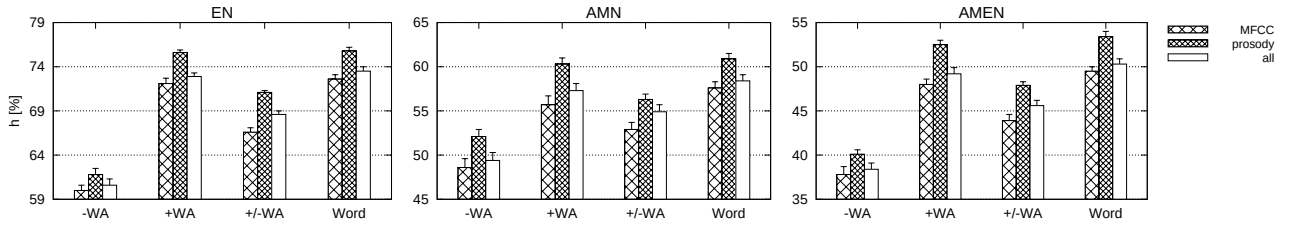

Figure 2: *Classification results per configuration (***EN***,* **AMN***,* **AMEN***) and per group of features: 97 prosody-based features (energy + pitch + zero-crossing-rate + harmonic-to-noise ratio + duration); 288 MFCC-based; and all features together (385). Performance figures, in %, are harmonic means (*h*) over the classes of weighted and unweighted recalls. Error bars represent standard errors.*

spectral features by exploiting the findings in previous studies (e. g. [8]). Features are calculated using the openEAR feature extractor [9]. The whole set comprises 16 low-level descriptors: *energy, pitch, zero-crossing-rate, harmonics-to-noise ratio, and the first 12 Mel-frequency cepstral coefficients*. To each of them, we add the delta coefficients computed to cope with feature dynamics. Next, 12 functionals are applied on a unit basis: *mean, standard deviation, kurtosis, skewness, minima and maxima with relative positions, range, and linear regression coefficients with the regression error*. *Duration* in frames, as obtained from the forced alignment, is also added to the feature set. The total feature vector per word or per syllable, depending on the analysis, contains up to 385 attributes (288 MFCC-based, 97 prosody-based).

The data are partitioned into three balanced splits (as in [8]) meeting the following requirements (in order of priority): speaker-independence, similar distribution of labels, balance between the two schools, and balance between genders. In order to have a more balanced distribution for training, we upsampled all classes but **N***eutral* to uniform distributions.

Classification results shown in Fig. 2 are obtained by 3-fold Cross-Validation (CV). As classifier we opted for Support Vector Machines (SVM) trained by Sequential Minimal Optimisation with linear kernel [10]. Parameters are optimized separately on the three different training sets also by CV.

Figures of classification performance are harmonic means (h) of weighted (*WA*) and unweighted averaged (*UA*) recalls: for very unbalanced datasets like ours this reduces the spread between *UA* and *WA*. In this way we give each class the same importance, and we neither penalize sparse classes such as **M***otherese* nor very frequent ones such as **N***eutral*. In Fig. 2 and 3 we scale the y-axis (h) differently as we are not interested in absolute performance but, mainly, in the relation between +WA, -WA, and Word across the three configurations; it could be expected that fewer classes yield higher performance.

In the following discussion we also consider the accuracy of the results: standard errors are estimated by running the $bootstrap$ algorithm [11], a Monte Carlo method: replications of h are obtained by random sampling with replacement from the training sets. Differences (in h) of at least two standard errors greater than zero correspond to p-values of $\simeq 5\%$. Finally, we draw conclusions on figures that show consistency across configurations and feature sets, and should therefore be less prone to multiplicity effects.

4. Discussion

The results displayed in Fig. 2 confirm our H1 formulated in Sec. 1. Throughout the three different constellations, independently of the feature set, -WA syllables perform worse than all other units; Word are the best along with accented syllables (+WA). The almost identical performance of +WA and Word could, to some extent, be expected as many words are made of one syllable only; but, as can be inferred from the histogram of Fig. 1, there is also a considerable number of words ($\simeq 30\%$) with two or more syllables.

These results, corroborated by the mediocre performance of the mixture +/-WA, do not mean that there is no information entailed in -WA syllables that, alone, perform well above chance level. However, in combination with +WA syllables when modelling words (Word), -WA syllables do not give any added value: the small improvements ($\leq 1.9\% \pm 1.2$ for MFCC, **AMN**) are not significant for any configuration. In other words, modelling words or only syllables carrying the word accent (lexical stress) is equivalent. Also note that differences among acoustic units +WA and -WA are larger for **EN** and **AMEN**, i. e. those constellations where **E***mphatic* is involved, and lower for **AMN** but still ample ($\geq 7.1\% \pm 1.4$ for MFCC).

The situation is very different if we concentrate on the word 'Aibo' only (2257 tokens). From Fig. 3 we notice that the behaviour of the acoustic units across configurations is quite similar; more specifically, the +WA syllable does *not* contribute more than the -WA syllable. This might be traced back to the special role of this word and its use in the interaction: 'Aibo' denotes the communication partner and is practically always used as vocative, i. e. produced with continuation rise, mostly without shortening and/or centralisation of the -WA syllable; cf. the notion of 'calling contour' [12], or 'vocative chant', i. e. H*!H% in ToBI notation. The duration plots in Fig. 4 clearly show this difference: for all 'Aibo' words there is no clear peak of -WA syllables [bo:], whereas the frequency peak for the syllables of all other words is sharper and at around 12 frames. Similar considerations hold for MFCC features alone (Fig. 3): with identical segmental information, the 2nd syllable can be modelled very consistently.

In Fig. 2 we display classification results using three differ-

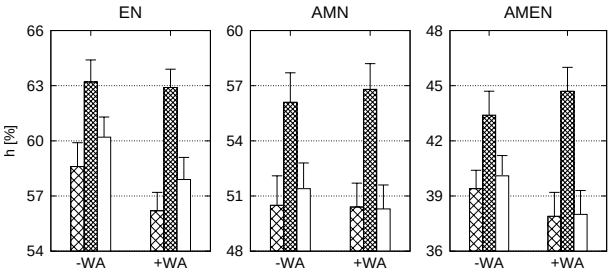

Figure 3: *Classification results using (-*WA *or* +WA*) syllables from 'Aibo' words only. Performance figures in* h[%]*. Error bars represent standard errors. Same legend as Fig. 2.*

ent subsets of features, namely prosody-based, MFCC-based, and all the features together. The idea is to isolate the implicit modelling of the words/syllables induced by MFCC features. This is probably even more crucial for our data, as they consist of many mono-syllabic words: the classifier could learn the syllables that mainly fall in one specific class rather than their acoustic form. On the other side, as shown in [8], MFCC are clearly useful in emotion recognition tasks. However, as can be seen from Fig. 2, acoustic information coded in MFCC does *not* add up to prosodic information. One possible explanation is that, for short units like syllables and short words, prosodic phenomena (like duration or pitch rise/fall) are sharper and more consistent than on longer units like chunks or turns.

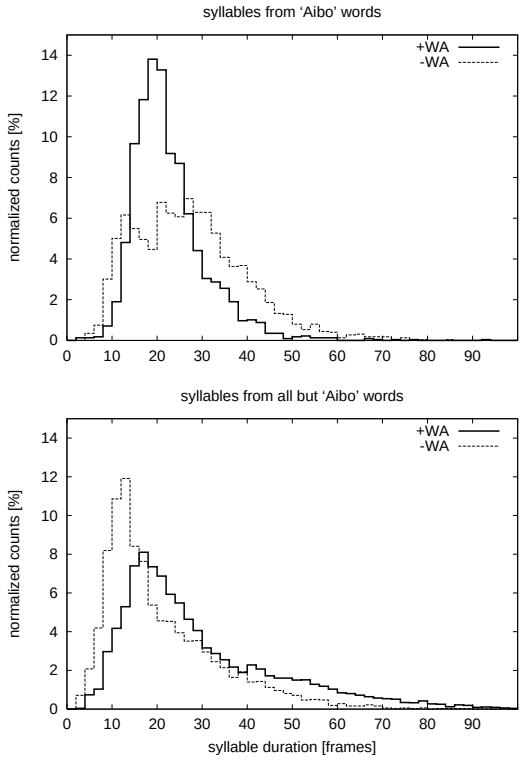

Figure 4: *Normalized histograms of syllable durations in frames (*1 *frame* = 10 *ms) for 'Aibo' words only and for all but 'Aibo' words, for both* +WA *and* -WA *syllable durations.*

5. Concluding remarks

There might be a simpler explanation for +WA being better than -WA: +WA syllables are generally longer, and this is probably the reason why 'more can happen' on these syllables, i. e. we can find more pronounced parameter values for +WA than for -WA. So we would not have to refer to emotion being – somehow – part of linguistics. On the other hand, there is of course linguistic reasons for +WA being longer and more pronounced; thus, it is not a typical chicken-and-egg problem where we do not know what came first.

Note that classification performance, full coverage of the data (open-microphone setting), or feature evaluation have not been the focus of this study. Moreover, we want to stress that in this study – the same way as in any other study using realistic data – only a subset of classes can be modelled which can be subsumed under 'appraisal of the interaction within a specific type of communication: giving commands to a pet robot'. This means that the indication of emotion is more conscious and thus, most likely, more according to linguistic structure than, e. g., when speaking in a thoroughly depressed or sad mood. Yet it has to be shown whether something like sadness really is modulated onto the speech chain in a fully global way, without taking into account linguistic structure and by that, stress patterns, at all.

As mentioned above, we counterfactually assign +WA even to normally unstressed articles or other function words, following the simplistic rule 'each mono-syllabic word carries word accent'. In our context, this means that our results are conservative because if we modelled unstressed mono-syllabic words as well, most likely the differences would have been even more pronounced. This holds for the word level. Apparently, words carrying the phrase accent are as well more pronounced than words that do not: this is well known from studies on (semantic) salience. The same way, such words might be 'emotionally prominent' and thus, better candidates for emotion classification than non-salient words.

6. Acknowledgements

This research has received funding from the European Community under grant No. RTN-CT-2006-035561 (S2S), No. IST-2001-37599 (PF-STAR), No. IST-2002-50742 (HUMAINE), and No. FP7-2007-211486 (SEMAINE). The responsibility lies with the authors.

7. References

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion Recognition from Speech: Putting ASR in the Loop," in *Proc. of ICASSP*, Taipei, 2009, pp. 4585–4588.

[2] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private emotions vs. social interaction — a data-driven approach towards analysing emotions in speech," *User Modeling and User-Adapted Interaction*, vol. 18, pp. 175–206, 2008.

[3] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin: Logos Verlag, 2009, PhD thesis.

[4] M. Shami and W. Verhelst, "Automatic Classification of Expressiveness in Speech: A Multi-corpus Study," in *Speaker Classification II*, C. Müller, Ed. Berlin: Springer, 2007, pp. 43–56.

[5] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-Forcing Hierarchical Functionals for Paralinguistics: a Waste of Feature Space?" in *Proc. of ICASSP*, Las Vegas, 2008, pp. 4501–4504.

[6] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. of ICME*, Amsterdam, 2005, pp. 474–477.

[7] K. Demuynck, J. Roelens, D. V. Compernolle, and P. Wambacq, "SPRAAK : an open source 'SPeech Recognition and Automatic Annotation Kit'," in *Proc. of Interspeech*, Brisbane, 2008, pp. 495–499.

[8] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals," in *Proc. of Interspeech*, Antwerp, 2007, pp. 2253–2256.

[9] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. of ACII*, Amsterdam, 2009, pp. 576–581.

[10] I. H. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[11] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

[12] K. Pike, *The intonation of American English*. Ann Arbor, Michigan: University of Michigan Press, 1945.