MOTHERS, ADULTS, CHILDREN, PETS — TOWARDS THE ACOUSTICS OF INTIMACY

Anton Batliner[1], *Björn Schuller*[2], *Sonja Schaeffler*[3], *Stefan Steidl*[1]

[1]Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
[2]Institute for Human-Machine Communication, Technische Universität München, Germany
[3]Speech Science Research Centre, Queen Margaret University, Edinburgh, GB
batliner@informatik.uni-erlangen.de

ABSTRACT

In this paper, we investigate acoustic features which differentiate the two speech registers *neutral* and *intimate* within different constellations of speakers and addressees. Three different types of speakers are considered: mothers addressing their own children or an unknown adult, women with no children addressing an imaginary child or an imaginary adult, and children addressing a pet robot using both intimate and neutral speech. We use a large, systematically generated feature vector, upsampling, and SVM and RF for learning. Results are reported for extensive test-runs facing speaker-independency and using PCA-SFFS vs. SVM-SFFS for feature ranking. Classification performance and most relevant feature types are discussed in detail.

Index Terms— Affective Speech, Intimacy, Emotion, Feature Selection.

1. INTRODUCTION

Para-linguistic variability within speech quite often serves the function of signalling membership, be this belonging to a certain regional group (dialects) or a certain social class (sociolects).Throughout, the *speech register* people use can be part of formal (symmetric or non-symmetric transactional) or intimate interaction, for instance between members of the same family. We want to address 'neutral' speech — i.e. speech that does not convey emotions and is not triggered by a certain bond with the interlocutor — and speech establishing a certain relationship between speakers and addressees; the register we are dealing with we want to call the register of *intimacy* which is, for example, typically produced when parents address their own children. Intimate adult-adult interaction between lovers can display similar characteristics.

Depending on the different personnel and their different tasks recorded for our databases, we speak about Child-Directed Speech (CDS) if adults address children, of Adult-Directed Speech (ADS) if adults address adults, and of Pet-Directed Speech (PDS) when referring to children addressing a pet robot (as in [1, 2]).

CDS denotes the speech register that speakers and in particular parents use to address a child of language acquisition age; it is characterised by raised pitch, wider pitch range, exaggerated prosody, hyper-articulation, slower speech rate, and reduced linguistic complexity, and is assumed to constitute an adaptation that permits mothers to control the child's arousal, and to elicit the child's attention [3]. Importantly, CDS is widely reported to be a speech style that is expressing emotions as well as being triggered by emotions. It has been suggested [4] that the prosodic characteristics of CDS resemble those associated with the vocal expression of positive affect, namely higher pitch, wider pitch range, as well as higher acoustic energy, with only speech rate being described as slower in CDS compared to speech expressing positive affect [5]. This observation is in line with views that communication with young infants is predominantly affective in nature. There is growing evidence that affective communication is crucial in early child development [6].

Less is known about CDS features of non-kin interlocutors who may not have an affective bond with the child. [7] found that pitch and pitch range increased when speaking to an imaginary child suggesting that CDS of non-kin interlocutors is similar to parental CDS. Pitch and pitch range increased even more when the adults were speaking to a real child. Features of CDS seem to be in fact also present in other speech registers, in order to signal a certain amount of intimacy or affect. PDS as well as so-called lovers' speech are such examples. In line with this, we could show in [1] that children addressing a pet robot dog utilise prosodic features similar to those found in CDS when expressing positive affect towards the toy. This analysis was based on an extensive set of acoustic parameters; comparisons with maternal CDS were made based on comparisons with findings from the relevant CDS literature.

In the present paper, we want to compare three different speaker groups in two different settings, namely in the *'neutral'* register and the *'intimate'* register: mothers (M) addressing their own child or another adult, 'non-mothers', i.e. adult (A) females without own children, addressing imaginary children or imaginary adults, and children (C) addressing a pet robot using either the neutral or the PDS-like register. As a common terminology for the three different constellations, we resort to 'neutral' vs. 'intimate' as two different speech registers.

2. MATERIAL AND ANNOTATION

2.1. MATERNAL CHILD DIRECTED SPEECH

A referential communication task, based on the study design introduced in [8], was set up with two conditions that were run as separate trials: CDS and ADS. The speaker's task comprised of giving instructions to a listener whose task it would be to manipulate soft toys laid out on an array according to the instructions. The set up of arrays and soft toys was counterbalanced across speakers. In each condition, speakers used eight pre-defined instruction sentences of the type *"Touch the frog with the spoon"*. The sentences were presented in a booklet and were the same for both conditions, but were counterbalanced across speakers depending on the array of soft toys assigned to the speaker. For mothers the addressee was their own child (CDS) and an unknown adult (ADS), respectively. The recordings were carried out in a sound-treated room at the Department of Psychology, University of Stirling. 24 English mothers, mean age 35 years (ranging from 23 to 46) and their infants (N=24), aged between 2;0 and 3;8, took part.

They were told that they were taking part in a pilot study aimed at exploring how children follow instructions and from which age they are able to do so. For each trial one experimenter was present who oversaw the recordings and instructed the participants as well as one confederate who acted as the adult addressee. Both the experimenter and the confederate were unknown to the mothers. The order of addressee was counterbalanced. Half of the mothers first addressed their child and then the adult confederate, the other half first addressed the adult confederate and then their child. When mothers addressed the adult, the child was taken to an adjacent room, so as to keep the child out of sight and to avoid dual-tasking or even child-directed speech.

2.2. IMAGINARY CHILD DIRECTED SPEECH

A referential communication task identical to the task for the MATERNAL CDS database was administered to women with no children of their own. For these non-mothers the addressee was an imaginary child (CDS) and an imaginary adult (ADS), respectively. The recordings were again carried out in a sound-treated room at the Department of Psychology, University of Stirling. 24 women, mean age 27 years (ranging from 21 to 42 years), all native speakers of English, were assigned as non-mothers. Non-mothers were told that they were recorded giving instructions that would later be used for another study to test a 'game' in which listeners would have to follow instructions. They were shown the arrays and told how the game works. The order of imaginary addressee (adult vs. child) was counterbalanced. Participants were only told about the second type of imaginary addressee after they had finished the recordings for the first type of imaginary addressee. This had proven to be less confusing for speakers in pilot runs. When asked to address an imaginary adult, participants were instructed to imagine speaking to a friend or acquaintance. When asked to address an imaginary child, participants were instructed to imagine speaking to a two- to three-year old they might know or have seen in the department's toddler group.

2.3. CHILDREN'S PET DIRECTED SPEECH

The database used is a German corpus with recordings of children communicating with Sony's Aibo pet robot [1, 2, 9]. The children were led to believe that the Aibo is responding to their commands but it 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, by that provoking emotional reactions. The data was collected at two different schools from 51 children (age 10 - 13 years; 21 male, 30 female; about 9.2 h of speech without larger pauses). Five labellers (advanced students of linguistics) listened to the turns in sequential order and annotated independently from each other each word as neutral (default), PDS or as belonging to one of 9 other classes. If three or more labelers agreed, the label was attributed to the word (majority voting MV); in parentheses, the number of cases with MV is given for the considered classes: PDS (1260), *neutral* (39169).

For the present study, we mapped the words onto syntactically/ semantically meaningful chunks [9]. We first selected *neutral* and PDS chunks (C_{all}), and in addition, only /Aibo/ tokens that are either *neutral* or PDS (C_{Aibo}) in order to keep the data as comparable as possible with the two other sets, and balanced this subset accordingly. For C_{Aibo}, we approached the conditions of balanced speaker numbers per split, and identical amounts of utterances of both classes per speaker. 21 of the speakers had at least one such turn in both intended classes. Using all their turns in class-balance led to a total of 220 turns of 21 speakers opposing the 192 (2 classes x 24 speakers each x 4 predefined sentences) and 24 for mothers and non-mothers, each. Three cross-fold splits have been constructed per group meeting the requirements of speaker-independence and balancing in terms of classes and speakers. For mothers and non-mothers this is accomplished straight-forwardly by separating them into groups of 8 subjects per split. For C_{Aibo}, this had to be balanced as turn-couple numbers vary between 2 and 28 per speaker. A splitting could be found that preserves school and gender balance as well as variance of samples per speaker while having 7 speakers within each split. For C_{all}, an according splitting strategy was chosen. As 586 PDS chunks oppose 1998 *neutral* chunks, it was decided upon 3x up-sampling of the first throughout classifier training to obtain balance.

3. FEATURE EXTRACTION

A strictly systematic generation of features was chosen for the construction of a large feature space as basis for subsequent selection of relevant features. Our basis is a set of 37 typical acoustic Low-Level-Descriptors (LLD) well known to carry information about paralinguistic effects shown in Tab. 1. We group the features into the common types duration (D), energy (E), pitch (P), formants (F), cepstral (C), and voice quality (V). Duration features thereby model temporal aspects having milliseconds (ms) as unit. Voice quality is covered by jitter and shimmer (micro-perturbations based on pitch and intensity, respectively) and Harmonics-to-Noise Ratio (HNR).

In order to calculate LLDs, first the speech signal is transformed to 16 kHz, 16 bit. In general, a Hamming window function is used, except for the calculation of F0 and HNR where a Hanning window has been chosen. We use 100 fps with semi-overlapping windows. Energy resembles simple log frame energy. F0 and HNR calculation base on the time-signal ACF with window correction. Formants base on 18-point LPC with root-solving and a pre-emphasis factor $\alpha = 0.7$. F0 and formant trajectories are globally optimized by use of Dynamic Programming. LLDs are smoothed by according techniques as semi-tone-interval filters or simple moving average low-pass-filtering to overcome noise. As a next step we add delta coefficients for each LLD. Following the typical static classification strategy used in the related recognition of emotion, we next employ a total of 19 statistical functionals to each of the 2x37 LLDs. The obtained multivariate time series of variable length is projected on a single 1406 dimensional feature vector. Here again we decided for a typical selection of common functionals covering the first four statistical moments, quartiles, extremes, ranges, positions, and zero-crossings as depicted in Table 1. The three position related functionals lead to a sub-group of features with the physical unit of ms which are treated as duration features, though having a number of diverse LLDs as basis. We refrained from inclusion of further duration related features such as those based on e.g. lengths of pauses or syllables, because this information cannot easily be integrated in the strictly systematic generation approach: it is modelled in a general value series rather than in a time series. The number of features per type is given in Tab. 3.

4. CLASSIFICATION

We computed a 3-fold cross-validation with Random Forests (RF) and Support Vector Machines (SVM) with polynomial kernel and pair-wise mulitclass discrimination. We report the F value as a unique performance measure; here, F is defined as the uniformly weighted

Low-Level-Descriptors (2x37)	Functionals (19)
(Delta) Pitch (F0)	Mean, Centroid, Std. Dev.
(Delta) Energy	Skewness, Kurtosis
(Delta) Envelope	Zero-Crossing-Rate
(Delta) Formant 1-5 Amplitude	Quartile 1,2,3
(Delta) Formant 1-5 Bandwidth	Quartile 1 - Minimum
(Delta) Formant 1-5 Frequency	Quartile 2 - Quartile 1
(Delta) MFCC Coefficient 1-16	Quartile 3 - Quartile 2
(Delta) HNR	Maximum - Quartile 3
(Delta) Jitter	Max., Min. Value, Range
(Delta) Shimmer	Relative Max., Min. Pos.
	Pos. 95% Roll-Off-Point

Table 1. Low-Level-Descriptors and functionals used throughout systematic construction of a large acoustic feature space

harmonic mean of accuracy and unweighted mean recall (note that the data is partly unbalanced.) Results reported employ speaker-normalization which is realized by a simple normalization of each feature by its mean and standard deviation for each speaker individually. Thereby the whole speaker context is used. This has to be seen as an upper benchmark for the ideal situation, where a speaker could be observed with a variety of utterances within the aimed at speaking styles. Yet, it is not necessary to know explicitly the according class for this purpose. Tab. 2 displays the classification results within the groups M, A, and C, and for M and A taken together, as well as across the groups, i.e., one group for training, the other one for testing. It can be seen that modelling M and A together results in some lower performance. Across groups, performance is at chance level if children are taken as either train or test sample (not shown in Tab. 2). Classification is best for the 'real' mothers M for both SVM and RF, and considerably lower for the 'fake' mothers A. Obviously, using only information pertaining to one word makes the task more difficult (C$_{Aibo}$) than using chunks (C$_{all}$) — although there are many one-word chunks such as /Aibo/, /stop/, etc. Note that the classification task for the children could be expected to be more difficult because the reference is perceptual labelling and not, as is the case for mothers and adults, a clear experimental setting; yet F_{RF} values for both C$_{Aibo}$ and C$_{all}$ are higher than those for the A group. This indicates that behaving 'as if' in an imaginary setting, that is acting, yields less pronounced characteristic traits than addressing ones own child or a pet-robot, respectively.

5. MOST IMPORTANT FEATURE TYPES

Two different strategies are used to find most relevant feature types: first, the feature space is transformed by Principal Component Analysis (PCA) in order to obtain a primary de-correlated representation. Next, the 50 PCs with according highest Eigen-value are chosen to actually reduce the dimensionality by conserving utmost covered variance. We decided for a fixed number of PCs over a cut-off criterion as ROC-curves to keep conditions constant throughout splits and data-sets. The number of 50 PCs resembles a reduction to 3.5% at this point and is a reasonable figure with respect to interpretability and computational effort for the subsequent secondary selection of best PCs. This second selection process employs the target classifier as wrapper-function in the closed-loop to optimize a highly compact set of PCs. Thereby the overall error is minimised for the latter classifier, as this is employed as optimization criterion. Likewise, instead of simply combining features of single high relevance, we rather find a further de-correlated complementary set. As search function, we use the popular Sequential Forward Floating Search (SFFS). In general, search functions do not lead to the overall optimum, even if employing a high number of back-steps in the case of SFFS. Thus another hard criterion was chosen, to keep conditions constant: 5 PCs were selected by SFFS, as a compromise between the observed optimum between 3 and 5 PCs throughout splits and sets in terms of maximum accuracy. As each PC still consists of a linear superposition of all original 1406 features, we have to cut off for a final time at this point to obtain an interpretable result. Due to the often very flat slope of the absolute contribution of features within PCs, we limited the maximum number for interpretation to 15; as second criterion within these 15 PCs, we took an Eigenvector value half the size of the maximum value. This strategy is applied three times for each possible pair of splits to overcome singular effects. We obtain up to 225 (5 PCs x 15 features x 3 split configurations) after back-transformation to the original feature space resembling a reduction to 16%. As a secondary strategy we directly search optimal feature sets in the original un-transformed feature space by use of SFFS. This resembles the more usual approach to feature selection in the field of emotion recognition. Yet it omits the first de-correlation step, which often clusters features of the same LLD or functional, thereby distorting the outcome if compression is very high. Also, final feature set sizes have usually to be kept larger without previous PCA. We decided in favour of the best 50 features for each pair of splits; a cut-off criterion would introduce variance on several layers, because ROC-curves are often overlaid with statistical noise or show very flat slopes. Generally, the wrapper classifier for SFFS thereby consequently is SVM, known as SVM-SFFS.

Train	Test	F_{RF}	F_{SVM}
M	M	76.6	78.6
A	A	70.3	74.5
M+A	M+A	68.7	65.6
C$_{Aibo}$	C$_{Aibo}$	71.4	64.2
C$_{all}$	C$_{all}$	72.8	71.1
M	A	68.8	65.1
A	M	72.4	73.4

Table 2. Classification Results for RF and SVM within groups (top) and across groups (below)

	D	E	P	F	C	V	D	E	P	F	C	V
#	222	64	32	480	512	96	222	64	32	480	512	96
	Mothers						Adults					
PC	16.8	2.5	5.4	40.6	26.8	6.2	15.6	1.8	6.2	17.4	59.0	0.0
SFFS	13.3	4.7	7.3	32.7	36.7	8.7	18.7	2.0	4.0	22.7	51.3	1.3
	C$_{Aibo}$						C$_{all}$					
PC	13.5	2.4	2.4	23.0	55.6	3.0	5.9	6.9	6.2	33.8	37.9	8.9
SFFS	17.3	3.3	3.3	23.3	47.3	5.3	6.7	5.3	6.7	29.3	51.3	0.7

Table 3. Summary of feature selection with feature types abbreviated as introduced in section 3; higher values denote higher impact; description in the text

Tab. 3 summarizes the distribution of features for M, A, C_{Aibo}, and C_{all} for the three splits and for PCs and SFFS across all three pairs of splits; the impact (the 'share') of each feature type is shown in percent. To give an example: the SFFS sum of D features for M was observed to amount to (11+4+5)=20; 20/150=13.3%. The values per line for each speaker group sum up to 100% modulo rounding errors.

Most of the time, PC and SFFS selections yield similar results. Note that of course, the values depend as well on the overall number of features per type; we thus have to be cautious if comparing across types but we can do that for the same type across groups.

	D	E	P	F	C	V
group	222	64	32	480	512	96
M	+	-	+	+	-	+
A	+	-	+	-	+	-
C$_{Aibo}$	+	-	-	-	+	+
C$_{all}$	-	+	+	+	+	-

Table 4. 'Quantization' of Tab. 3: '+' meaning 'more important than in the other groups' (-)

.

In Tab. 4 we try to binarize the information given in Tab. 3: '+' means that in these groups, this feature type has been exploited to a higher extent than in the groups with '-'. Interestingly, there is no indication that either the language (English vs. German) or the age group (adults vs. children) are decisive factors: the signs span always across languages or age groups. (Note that this holds even if one of the binary differences would turn out not to be significant.) Factors that could be important are: same/different segmental structure, real mother-child interaction vs. imagined/pet-directed, and maybe even prompted (in the case of M and A) vs. non-prompted speech. Duration (D) seems to be more important, if segmental structure is kept constant (higher values for {M, A, C_{Aibo}}, lower ones for C$_{all}$). The opposite holds for energy (E). Pitch seems to be less important if it is the same short word with identical segmental structure (C_{Aibo}). Formants (F) are foremost used by M, second comes C$_{all}$ with its diverse segmental structure. MFCCs (C) are less exploited by M than by the other groups. Voice quality features (V) are most exploited by M, second comes C$_{Aibo}$. The two differences between M and A are thus a higher use of formants (F) and voice quality features (V) for M vs. a higher use of MFCC features (C), and rather no use of voice quality features (V) for A. MFCC features are modelling the spectrum in a more coarse-grained way than formants; accordingly there seems to be a sort of trading relation between formants (F) and MFCCs (C): if the share of the one is higher, the share of the other is lower. The average of F and C shares are close together for all four groups and for PC and SFFS alike: lowest for M (PC: 33.7%) to highest for C$_{all}$ (SFFS: 40.3%). It is more difficult to assess the role of feature types for the children, in comparison to M and A, because of the difference in segmental structure: C_{Aibo} is even more uniform than the sentences of M and A, and C$_{all}$ is free speech. However, it seems more likely that differences are rather due to these segmental factors than to a principled difference between CDS and PDS. Tab. 4 gives no indication that prompted vs. non-prompted is mirrored in the use of feature types because for all types, same signs can be found for both prompted and non-prompted speech.

Whereas for C$_{Aibo}$, only 10.5% of the features which represent our PCs are delta features, for C$_{all}$, it is 41.2%; for the SFFS features, it is 21.3% for C$_{Aibo}$, and 38.0% for C$_{all}$; M and A are in between C$_{Aibo}$ and C$_{all}$. This might indicate that larger units with segmentally different structure are better modelled with delta features, in contrast to short units such as our /Aibo/ words, or to units with partly identical segmental structure (M and A).

6. CONCLUDING REMARKS

The aim of this paper has been a comparison of four different constellations of speaker-addressee in two different speech registers: *neutral* and *intimate*. Within a state-of-the-art brute-force approach, we used a large, systematically generated feature vector for automatic classification and for subsequent interpretation of 'most important' types of features. Classification performance for this 2-class problem was between 70% and 80%, highest for the 'real' mothers, lower for the 'fake' mothers. We have seen that differences in language (English vs. German) or age-group (adults vs. children) do not seem to be relevant. The higher use of formants for M might indicate the well-known tendency towards hyper-articulation for CDS.

7. ACKNOWLEDGEMENTS

This work was partly funded by the EU in the projects PF-STAR under grant IST-2001-37599 and HUMAINE under grant IST-2002-50742. The responsibilty lies with the authors. S. Schaeffler has been funded through a Faculty Studentship by the University of Stirling, UK. Many thanks to Vera Kempe who has greatly contributed to the design and planning of the maternal and imaginary CDS studies.

8. REFERENCES

[1] A. Batliner, S. Biersack, and S. Steidl, "The Prosody of Pet Robot Directed Speech: Evidence from Children," in *Proc. of Speech Prosody 2006*, Dresden, 2006, pp. 1–4.

[2] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards More Reality in the Recognition of Emotional Speech," in *Proc. ICASSP 2007*, Honolulu, Hawaii, 2007, vol. IV, pp. 941–944.

[3] A. Fernald, "Human Maternal Vocalizations to Infants as Biologically Relevant Signals: An Evolutionary Perspective," in *Language Acquisition: Core Readings*, P. Bloom, Ed., pp. 51–94. Cambridge, MA: MIT Press, 1994.

[4] L. Singh, J. Morgan, and C. Best, "Infants' Listening Preferences: Baby Talk or Happy Talk?," *Infancy*, vol. 3, pp. 365–394, 2002.

[5] K. Scherer, "Vocal Communication of Emotion: A Review of Research Paradigms," *Speech Communication*, vol. 40, pp. 227–256, 2003.

[6] J. Locke, "First Communication: The Emergence of Vocal Relationships," *Social Development*, vol. 10, no. 3, pp. 294–308, 2001.

[7] J. Jacobson, D. Boersma, R. Fields, and K. Olson, "Paralinguistic Features of Adult Speech to Infants and Small Children," *Child Development*, vol. 54, pp. 436–442, 1983.

[8] J. Snedeker and J. Trueswell, "Using Prosody to Avoid Ambiguity: Effects of Speaker Awareness and Referential Context," *Journal of Memory and Language*, vol. 48, pp. 103–130, 2003.

[9] 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. INTERSPEECH 2007*, Antwerp, Belgium, 2007, pp. 2253–2256.