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 Center, Queen Margaret University Edinburgh, Great Britain

Introduction

- differentiation of the 2 speech registers *neutral* and *intimate*
- 3 different types of speakers:
  - mothers addressing their own children or an unknown adult
  - women without own children addressing an imaginary child or adult
  - children addressing a pet robot using both intimate and neutral speech

Data

1. Mothers addressing their own child or an unknown adult

- study of the Department of Psychology, University of Stirling:
  - How do children follow instructions?
  - From which age on are they able to do so?
- set of 6 instructions, partly ambiguous
  - *Touch the horse with the spoon* (non-ambiguous)
  - *Touch the fish with the flower* (non-ambiguous)
  - *Touch the dog with the flower* (ambiguous)
- addressees:
  - child-directed: their own child (at the age of 2;0 - 3;8)
  - adult-directed: unknown adult
- 24 mothers (23 - 46 years old, ⌀ 35 years)
- 192 recordings: 24 mothers · 4 instructions · 2 addressees
- language: English

2. Non-mothers addressing an imaginary child or an imaginary adult

- parallel communication task with 24 non-mothers (age: 21 - 42, ⌀ 27)
- addresses:
  - child-directed: imaginary child at the age of 2 - 3
  - adult-directed: imaginary adult (friend or acquaintance)

3. Children interacting with the Sony robot Aibo

- subset of the AIBO Emotion Corpus (University of Erlangen-Nuremberg)
- 21 children (10 - 13 years old)
- language: German
- level of analysis:
  - chunk level: 586 chunks *motherese*, 1998 chunks *neutral*
  - word level (only the word *Aibo*): 220 words

Features and Classification

Features

- typical acoustic low-level descriptors (LLD) on frame level
- applying functionals
- feature vectors on word, chunk, or file level

low-level-descriptors (2 · 37)	functionals (19)
(Δ) F_0	mean, std. dev., centroid
(Δ) energy	skewness, kurtosis
(Δ) envelope of the amplitude	zero-crossing-rate
(Δ) formants 1-5: frequency	quartile 1-3
(Δ) formants 1-5: amplitude	quartile 1 - minimum
(Δ) formants 1-5: bandwidth	quartile 2 - quartile 1
(Δ) MFCC 1-16	quartile 3 - quartile 2
(Δ) HNR	maximum - quartile 3
(Δ) jitter	max., min, range
(Δ) shimmer	position of rel. max./min.
	pos. 95 % roll-off-point

- 1,406 features: 2 · 37 LLD · 19 functionals

Classification

- 2-class problem: *neutral* vs. *intimate*
- 2 classifiers:
  - SVM with linear kernel
  - random forrests (RF)
- 3-fold cross-validation
- speaker independent
- normalization per speaker using the whole speaker context
- no explicit feature selection
- evaluation: F measure

Results

Train	Test	F measure RF	F measure SVM
Mothers	**M**others	76.6 %	**78.6 %**
Non-mothers	**N**on-mothers	70.3 %	74.5 %
Non-mothers	**M**others	72.4 %	73.4 %
Mothers	**N**on-mothers	68.8 %	65.1 %
M+N	**M+N**	68.7 %	65.6 %
Aibo (chunks)	Aibo (chunks)	72.8 %	71.1 %
Aibo (word)	Aibo (word)	71.4 %	64.2 %

- intimacy more pronounced for **M**others than for **N**on-mothers
- in accordance with subjective impression when listening to the data
- accuracy for children in the same range as for **N**on-mothers

Most Important Feature Types

Feature Types

- **D**uration (222)
- **E**nergy (64)
- **F_0** (43)
- **F**ormants (480)
- **MFCC** (512)
- **V**oice **Q**uality (96)

Two Ways of Feature Selection

1. SVM-SFFS
   - 50 features per split (3-fold cross-validation)
   - reduction to < 11 %
2. approach using eigenvectors
   - 50 eigenvectors with the highest eigenvalues
   - reduction to 5 eigenvalues using SVM-SFFS
   - selection of 15 “original” features per eigenvector
   - reduction to < 16 %

Surviving Feature Types

		Duration	Energy	F_0	Formants	MFCC	VQ
all	#	222	64	32	480	512	96
	%	15.8	4.6	2.3	34.1	26.4	6.8
feature selection	**M**others	15.1	3.6	6.4	36.7	31.8	7.5
	Non-mothers	17.2	1.9	5.1	20.1	55.2	0.7
	children (**C**hunks)	6.6	6.1	6.5	31.6	44.6	4.8
	children (**W**ord)	15.4	2.9	2.9	23.2	51.5	4.2
	⌀	13.5	3.6	5.2	27.9	45.7	4.3

feature type is more important for this data set than for others

feature type is less important for this data set than for others

Discussion

- no indication that age group or language are decisive factors
- segmental structure:
  - higher impact of duration if segmental structure is constant (**M**, **N**, **W**)
  - higher impact of energy and Δ features if segmental structure is variable (**C**)
  - F_0 less important if focus only on one single short word (**W**)
- degree of intimacy: higher impact of formants (**M**) vs. higher impact of MFCC features (**N**)

Friedrich-Alexander-Universität Erlangen-Nürnberg · Queen Margaret University Edinburgh · SSRC Speech Science Research Centre