Preparation of Verbal and Nonverbal Information of Speech in Spontaneous Conversation Database

Klara Vicsi

Budapest University of Technology and Economics,
Dept. forTelecommunications and Telemetics,
Laboratory of Speech Acoustics

http://alpha.ttt.bme.hu/speech/
vicsi@tmit.bme.hu

Human interaction

verbal chanel conveys
semantic content of speech

non verbal chanel convey
emotional state of the speaker
image conent of the massage

Range and number of emotions

complex and variable expressions of emotions are exist in different languages

can bring

variabilities into acoustical features

Major affect states:

In the frame of an EU project Emotion Research Group in Genova collected ~ 130 affected states in 5 languages

English	Deutsch	Français	Italiano	Espaňol
affectionate	Zuneigung	affectueux	affetuoso	afectuoso
afraid	verängsigt	apeure	impaurito	atemorizado
agitated	aufgeregt	agité	agitato	agitado
amazed	verwundert	stupéfait	stupefatto	estupefacto
amused	belustig	amusé	divertito	divertido
angry	ängerlich	en colére	arrabbiato	enfadado
annoyed	verärgert	agacé	seccato	fastidiado
calm	ruhig	calme	calmo	calmado
confused	verwirrt	confus	cunfuso	confundido
Depressed	deprimiert	déprimé	depresso	deprimido
. . .				

Labels describing affective states in five major languages. In K. R. Scherer (Ed.) (1988).

Hillsdale, NJ Erlbaum:*Facets of emotion: Recent research*

Problem with the selection of emotions:

It was impossible to find labels with exactly equivalent meanings across all languages.

It calls the very idea of a small number of universal basic emotions, at least with respect to the conceptualization of emotion in language.

These **commonly used emotions** are

happiness,
sadness,
anger,
surprise,
scorn/disgust

in psychology, linguistics and speech technology, and also described in the MPEG-4 standard.

In the last years:
a lot of speech emotional databases were constructed imitating these emotions mostly by artists

Problem with the imitated speech databases:

in the application of speech technology, real world data processing is necessary!!!

In the real conversation

these 5 emotion categories do not mask the emotions

and

the real world data differ much from acted speech

Real world data:

Main emotion categories were selected from everyday interaction
on the base of their frequency of usage in the Belfast natural audio-visual database.

PHYSTA 2001

television programmes: chat shows and religious programmes
studiorecordings of one to one interaction
(298 clips, 1clip 10-60 s in lengs)

Main emotion categories used in the Belfast natural database and their frequency of use (as first choice)

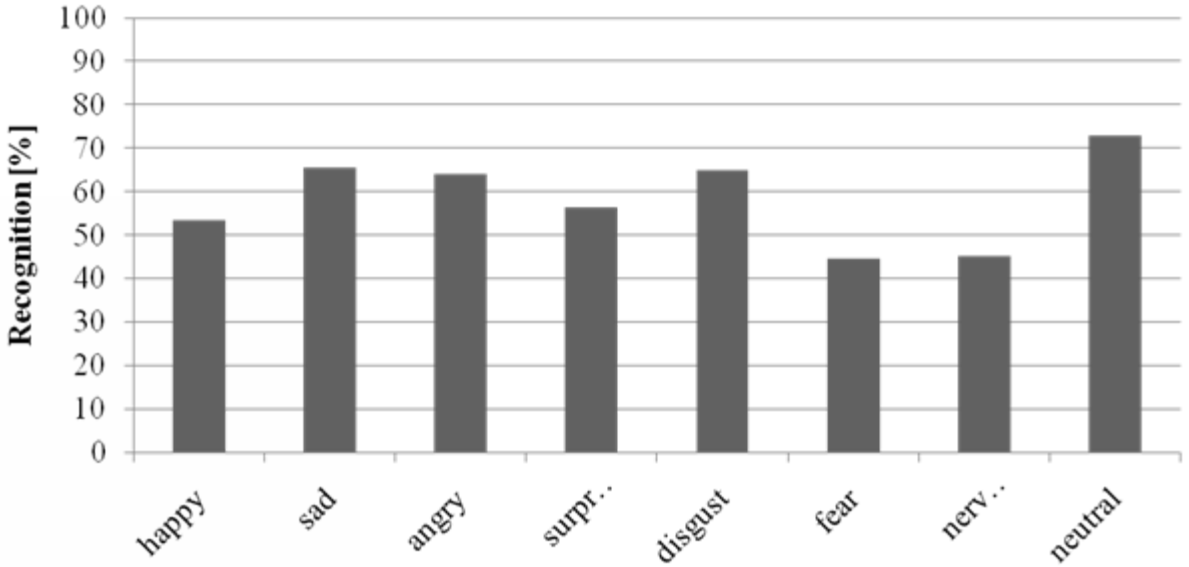
Label	Frequency of use	Broad group
Neutral	273	Not strongly motional
Angry	114	Strong negative
Sad	94	Strong negative
Pleased	44	Unoriented positive
Happy	37	Unoriented positive
Amused	26	Unoriented positive
Worried	19	Strong negative
Disappointed	17	Not strongly emotional
Excited	17	Oriented positive
Afraid	13	Strong negative
Confident	13	Not strongly emotional
Interested	12	Not strongly emotional
Affectionate	10	Oriented pozitív
Content	4	Not strongly emotional
Loving	3	Oriented positive

Problem with the selection of emotions:

The frequency of usage of emotions is different in different selection of databases

Further Problem:

Without linguistic context, the human speech emotion recognition is not better than 60-70%, according to our speech emotion perception experiment – where each sentences were articulated with different emotional meanings.



Speech emotion perception test results,
listened to a sequence of emotional sentences with the same linguistic content

(Sz.Tóth, D.Sztahó, K.Vicsi, 2007COST2102 Patras)

The semantic content (verbal channel) and

the general feeling and emotional state of the speaker (the non-verbal channel)

are expressed at the same time in speech,

and the semantic content contribute to the emotion recognition of human.

Presumably, if one want to get better recognition results by machine, the linguistic content and the emotional meaning must be recognised too.

Thus for emotion recognition the linguistic contents and its emotion must be annotated parallel in the speech database.

Emotion recognition through telephone line

It seems clear that vocal signs of emotion form only a part of a **multi-modal signalling** system.

Telephone conversation is an exception that proves the rule, because people adjust to the loss of other modes by adopting a distinctive _phone voice_ (Douglas-Cowieand Cowie, 1998).

It makes sense to study emotion in telephone conversations as **a purely vocal phenomenon.**

Our task in the Laboratory of Speech Acoustics:

The detection of the emotional state of the customer automatically,

through a telephone conversation with the dispatcher.

Our system description

During a speech conversation, especially if it is a long one, the speaker's emotional states are changing.

If we want to follow the states of the speakers we have to divide the continuous speech flow into segments, and thus we can examine how the emotional state of a speaker change segment by segment through the conversation.

Not only the acoustical parameters of emotions were examined and classified,

but

word and word connection statistics of different emotional text is planned to prepare, and word spotting, as well.

Our basic material is real world speech database:

Spontaneous everyday conversations between telephone dispatchers and customers, through telephone line were recorded, (by 8000 Hz and 16 bit sampling rate.)

The size of the database: 1040 calls were recorded, all together 60 hours speech conversation

The duration of one call changed between 1-30 minutes

Annotation:

the linguistic content (verbal channel) and

emotion state of the speaker (nonverbal channel) were processed parallel

The PRAAT tool was used for the segmentation and labelling while it is appropriate for parallel processing.

- The linguistic contents,
- the boundary of clauses,
- the emotion,
- the speakers with gender

were marked in the acoustical signal.

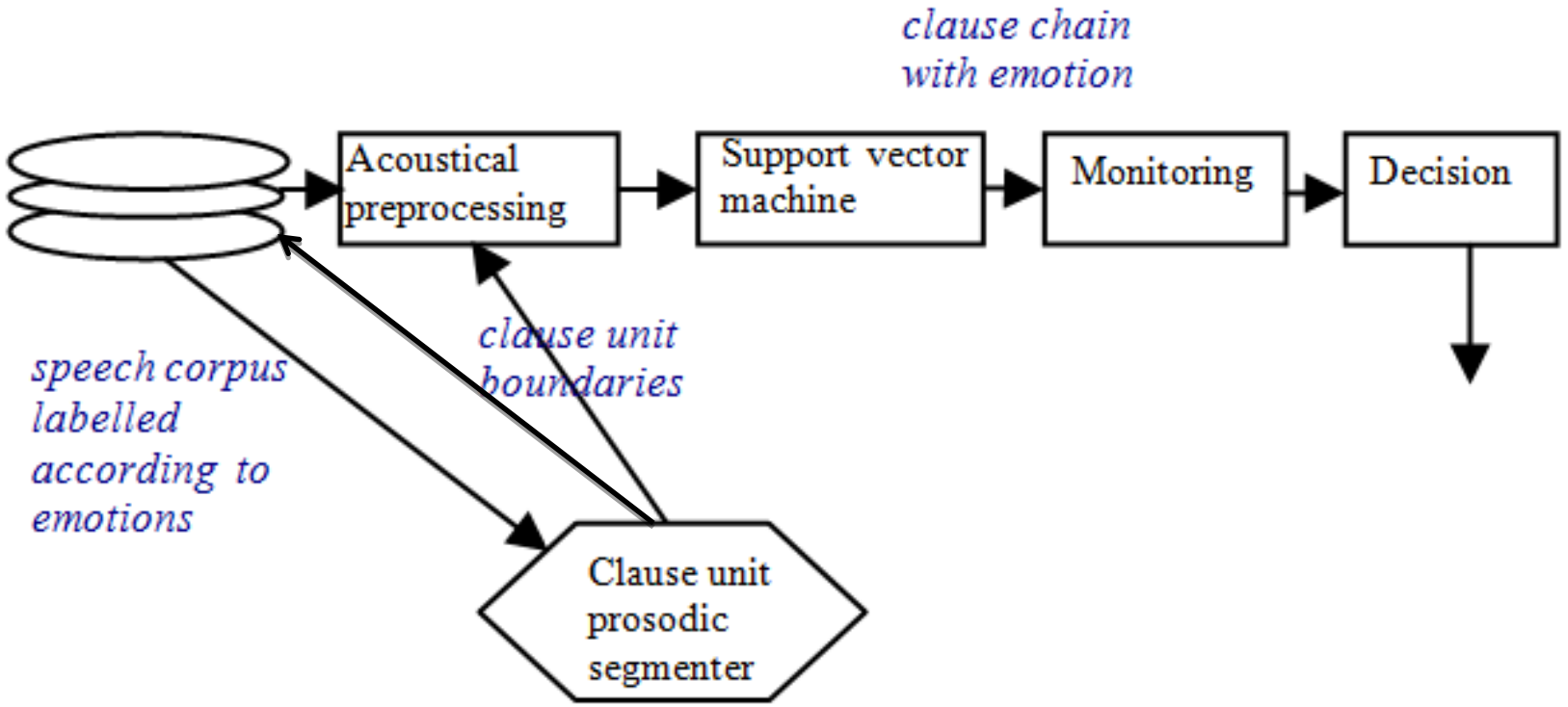
Four different emotional states were differentiated in the recorded dialogues

neutral (N), nervous (I), querulous (P), and others (E).

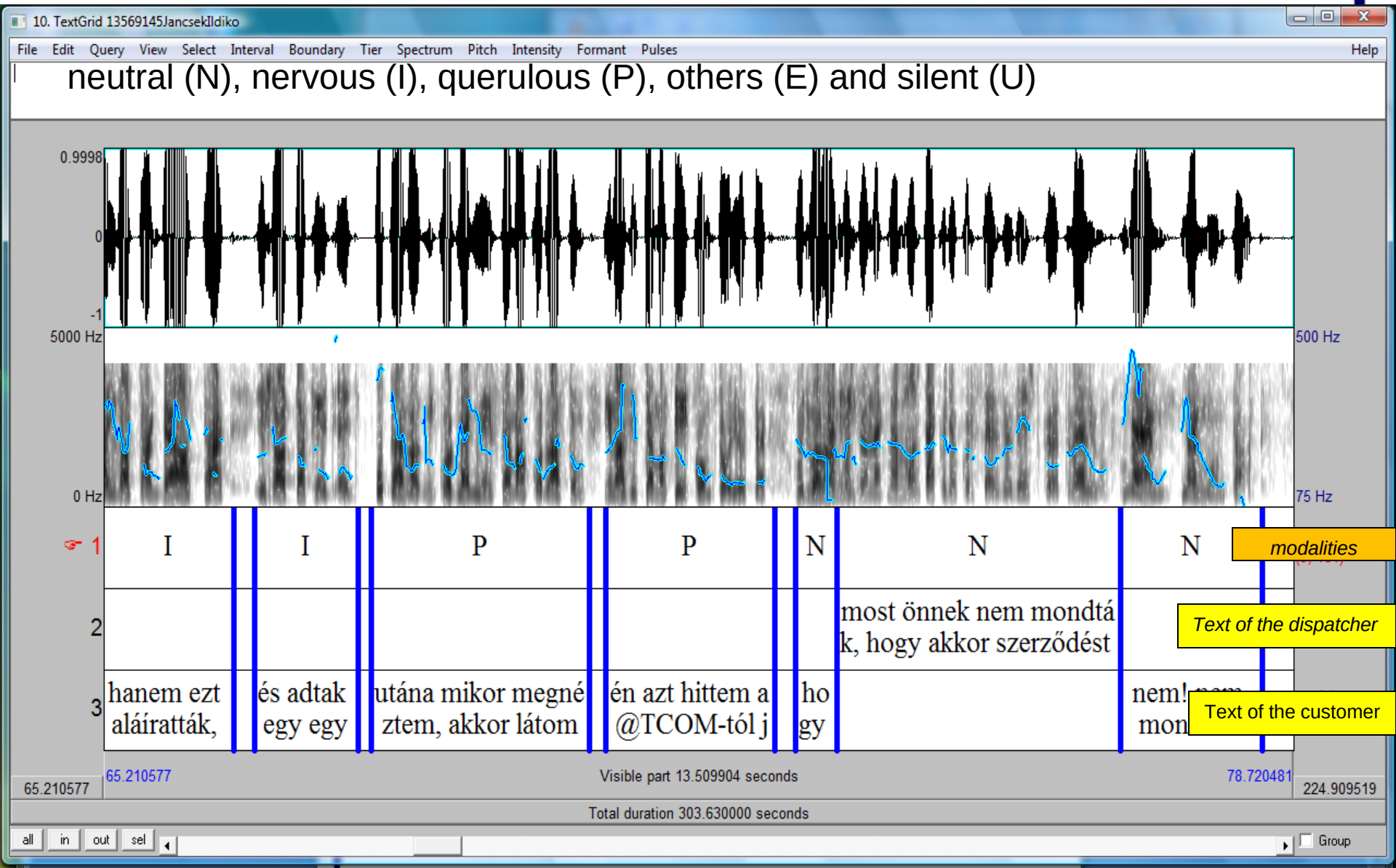
Practically there was no more emotion type in the 1000 calls

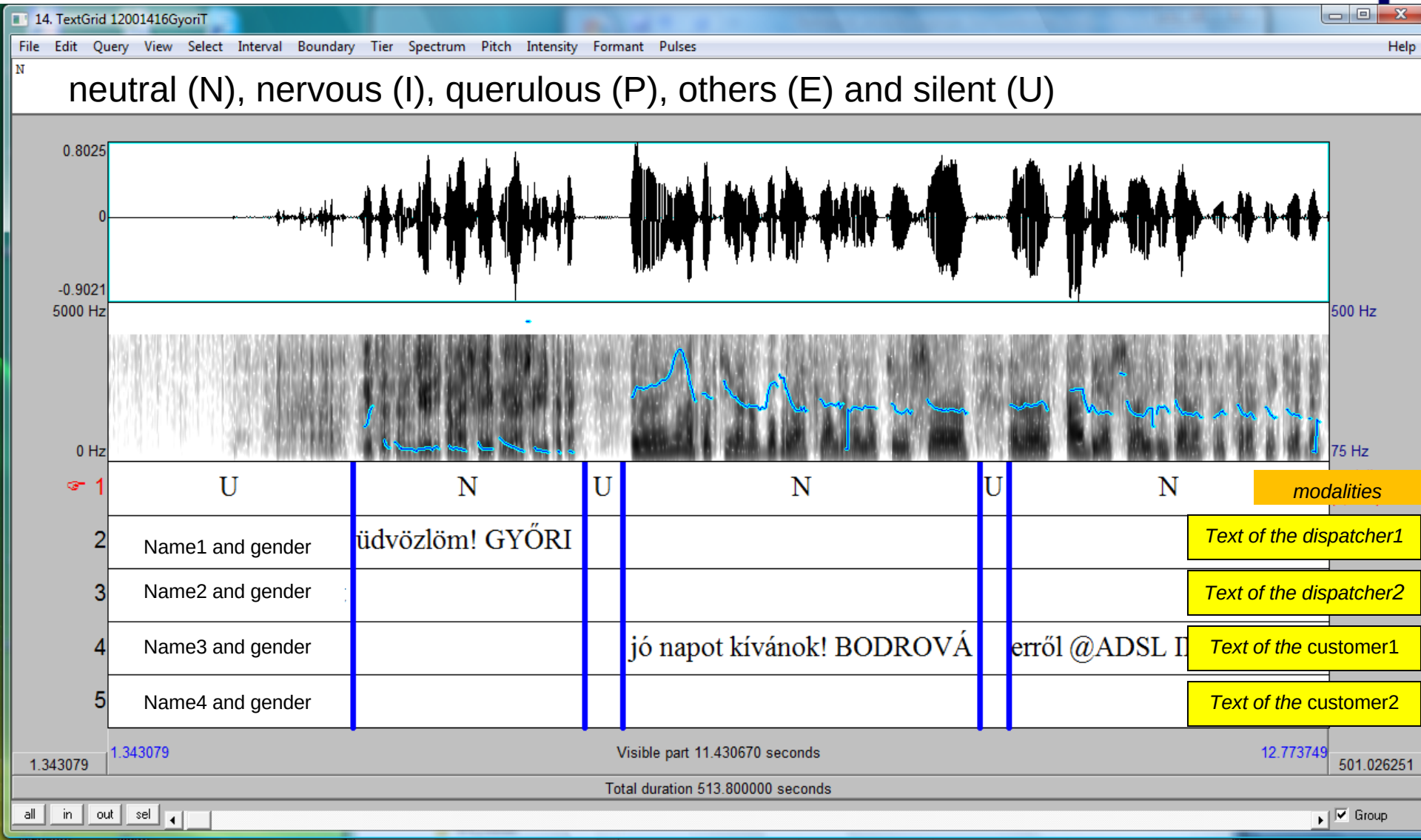
346 nervous clauses,
603 querulous,
225 others and
603 typical neutral clauses were selected from the neutral ones for the classification experiment.

The clause was selected as a segmentation unit in our system, on the base of the experiences of our earlier study.



Training and testing of our speech emotion classifier


10. TextGrid 13569145JancsekIldiko
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
neutral (N), nervous (I), querulous (P), others (E) and silent (U)
0.9998
0
-1
5000 Hz
0 Hz
500 Hz
75 Hz
1
I I P P N N N
2
most önnek nem mondtá k, hogy akkor szerződést
3
hanem ezt aláíratták,
és adtak egy egy
utána mikor megné ztem, akkor látom
én azt hittem a @TCOM-tól j
ho gy
nem! mon
modalities
Text of the dispatcher
Text of the customer
65.210577
Visible part 13.509904 seconds
78.720481
65.210577
224.909519
Total duration 303.630000 seconds
all in out sel
Group


14. TextGrid 12001416GyoriT
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
N
neutral (N), nervous (I), querulous (P), others (E) and silent (U)
0.8025
0
-0.9021
5000 Hz
0 Hz
500 Hz
75 Hz
1 U N U N U N
modalities
2 Name1 and gender
üdvözlöm! GYŐRI
Text of the dispatcher1
3 Name2 and gender
Text of the dispatcher2
4 Name3 and gender
jó napot kívánok! BODROVÁ
erről @ADSL I
Text of the customer1
5 Name4 and gender
Text of the customer2
1.343079
1.343079
Visible part 11.430670 seconds
12.773749
501.026251
Total duration 513.800000 seconds
all in out sel
Group
COST2102
Budapest
Dublin
emotion

For the training and testing

of our emotion classifier

the so called leave-one-out cross-validation (LOOCV) was used.

The error matrix in case of the four emotions is to be seen on Table

	E	I	N	P	Correct
E	49	26	62	88	22%
I	9	153	60	124	44%
N	14	38	398	153	66%
P	11	70	157	365	60%
				average	54%

***E, I, P, N** emotions are classified into separate classes*

The nervous (I), and querulous (P) emotions are hardly differentiated not only by the classifier, but by the humans too.

Thus I and P and E classes were closed up into one class, denominated as the “discontent” emotion.

Thus emotions are classified into two separate classes as discontent and neutral

	EIP	N	Correct
EIP	887	287	76%
N	335	839	71%
		average	73%

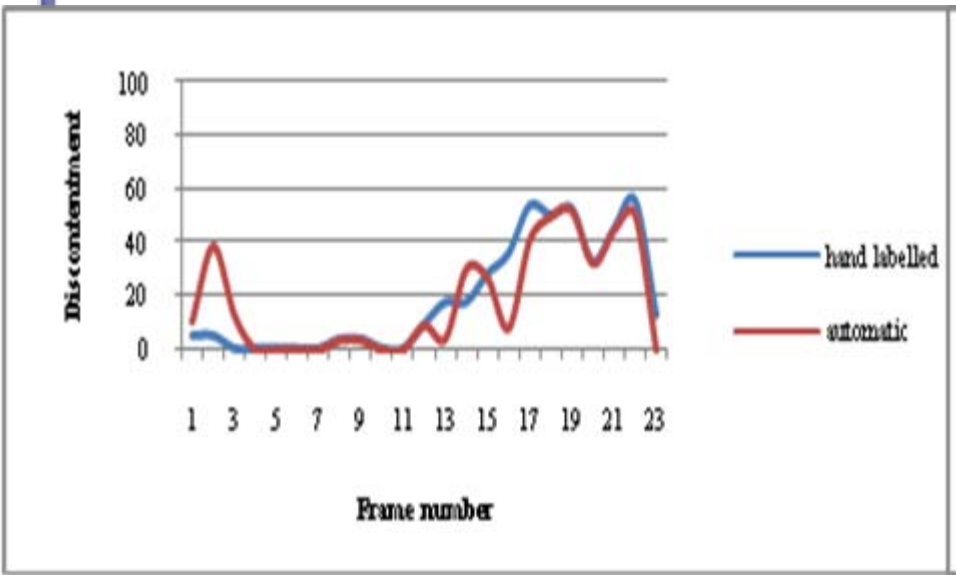
The aim:

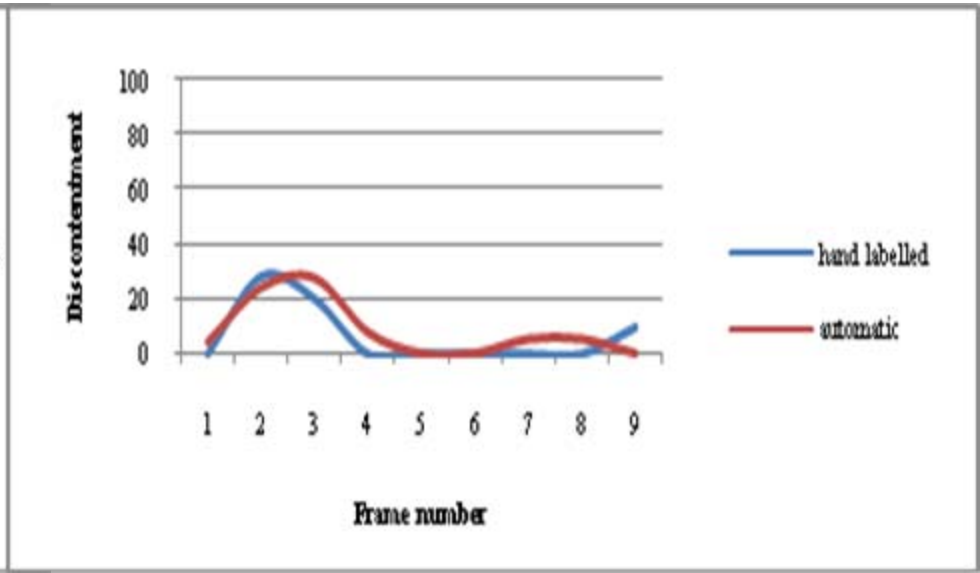
to detect the emotional state of the customer automatically, through a conversation.

With this object a 15 sec long **monitoring window** was selected, and the clauses, which were automatically classified as “discontent” emotion, were counted in this window.

Then, the window was moved forward, by 10 sec time steps.

The detection of the discontentment of the customer through the conversation:





The average deviation between the handmade and automatic labelling was 11,3 %.

On the base of this monitoring technique it is possible to sign when the discontentment level has reached a critical **“alarm level”.**

The average alarm detection error was 10,4% at the selected 30 % of discontentment,

but the reason of this error, in most cases, is only a small shift between the data of automatic recognition and hand labelling, resulting only a delay or foregoing of an alarm period.

Conclusions

- On the base of the acoustic parameters the two emotions: the discontent and neutral were classified with the error rate of 11,3%.
- The linguistic emotional content processing is under development
- Expectedly together with the linguistic processing the recognition increases over 90%.
- In case of real world data processing instead of 130 affected states we have differentiated only two stats with acceptable exactness,
- And the discontentness does not belong to the main 5 standardised category.

Thank you!

Problem with the size and the segmentation accuracy:

Different databases fit to speech synthesis and speech recognition.

- For speech syntheses some typical examples of the different emotions are collected.
- For speech recognition all the possible variations of emotions must be present in the database.

and

- for speech syntheses accurate segmentation is accepted,
- for recognition the accuracy of the segmentation is not so important.