

# Statistical analysis of Skype conversations: *recognizing individuals by their chatting style*



Candidato:  
Cristina Segalin

Relatore:  
Dr. Marco Cristani





# Abstract

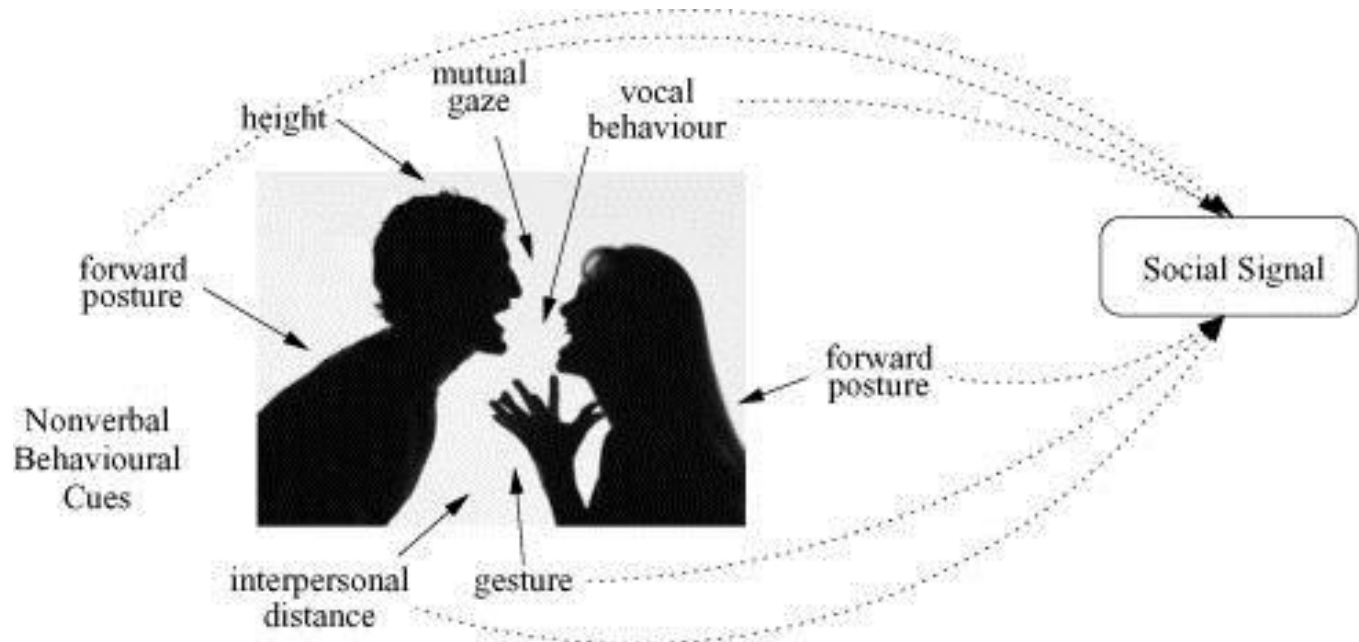
- **Goal of the thesis:** Recognizing the identity of a person by considering the way she/he chats
- **Facts:**
  - Each person has a particular style of writing
  - There are features which capture the style of writing
- **Assumptions:**
  - A chat is similar to a spoken conversation
- **Our contribute:**
  - Designing new stylistic features for analyzing chats, which capture the multimedia nature (written text + oral conversation) of a chat

- Introduction
- Our approach
- Experiments
- Conclusions



# Intro

- SSP: couples social psychology and Pattern Rec. aimed at modeling different behavioral aspects of a person.
- Conversational analysis (CA): describes the style of oral conversations. Conversational analysis is a subfield.





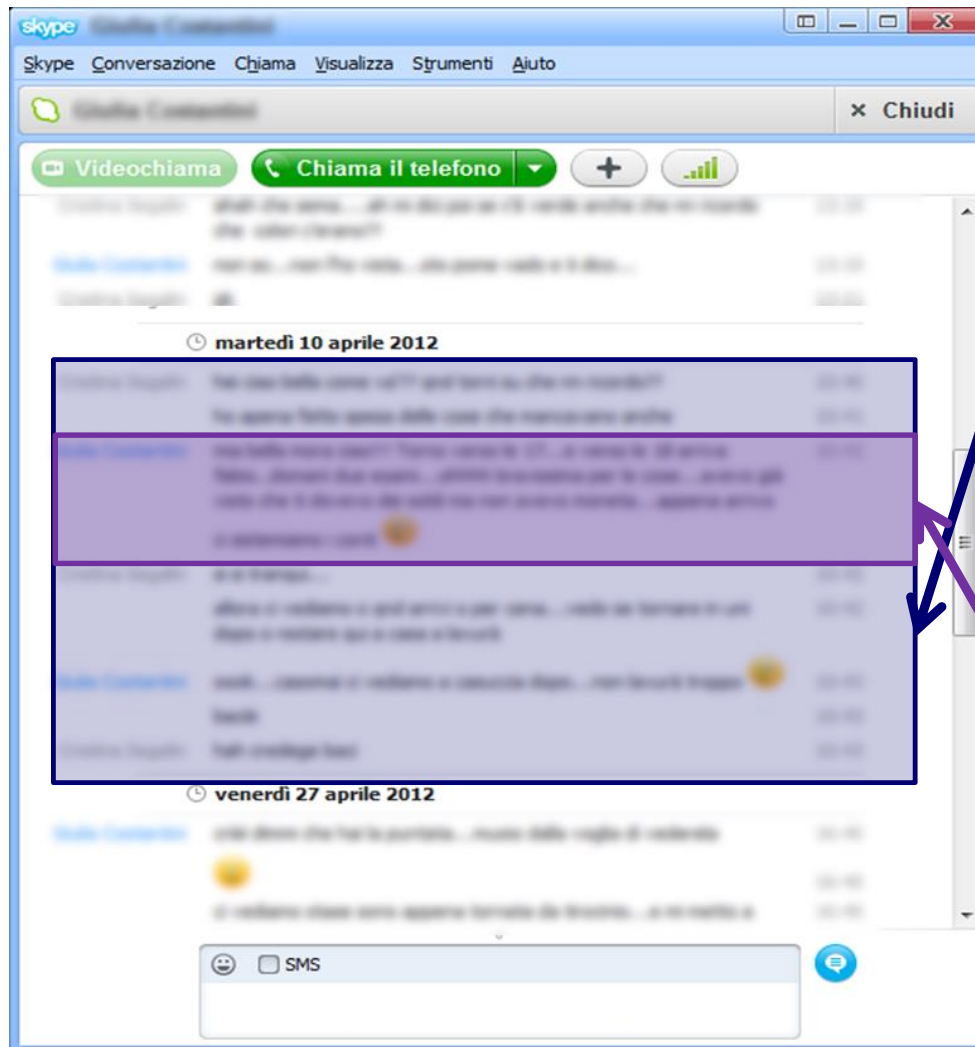
# Intro

- Stylometry is defined as statistical analysis of writing style. It is used to identify author of literary work, apply to music and fine-art paintings.
- Authorship attribution is the process of examining the characteristics of a piece of writing, an ancient text, a program code or comments on website to draw conclusions on its authorship.
- AA: old, traditionally on books, then on WEB, now on chats
- Based in the concept of style, subsumed by a set of stylometric features

# Tassonomy

| Group            | Description              | Examples   |
|------------------|--------------------------|--|
| Lexical          | Word level               | Total number of words (=M), # short words/M, # chars in words/C, # different words, chars per word, freq. of stop words                          |
|                  | Character level          | Total number of characters (chars) (=C), # uppercase chars/C, # lowercase chars/C, # digit chars/C, freq. of letters, freq. of special chars     |
|                  | Character—Digit n-grams  | Count of letter—digit n-gram (a, at, ath, 1 , 12 , 123)  |
|                  | Word-length distribution | Histograms, average word length  |
|                  | Vocabulary richness      | Hapax legomena, dislegomena  |
| Syntactic        | Function words           | Frequency of function words (of, for, to )   |
|                  | Punctuation              | Occurrence of punctuation marks (!, ?, : ), multiple !—?   |
|                  | Emoticons—Acronym        | :-), L8R, Msg, :( , LOL  |
| Structural       | Message level            | Has greetings, farewell, signature   |
| Content-specific | Word n-grams             | Bags of word, agreement (ok, yeah, wow), discourse markers—onomatopoe (ohh), # stop words, # abbreviations , gender—age-based words, slang words |
| Idiosyncratic    | Misspelled word          | Belveier instead of believer   |

# Feature extraction

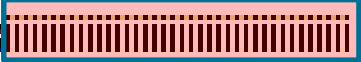





**AA standard:**  
consider the text as a whole

## Our ideas:

- consider the turn as atomic entity
- characterizing the turn taking via novel features

# Feature extraction

|                  |            |  |   |       |
|------------------|------------|--|---|-------|
| Cristina Segalin | s          |  |  | 11:56 |
| Cristina Segalin | dillo a me |  |  | 08:44 |

## Conversational features

Turn duration

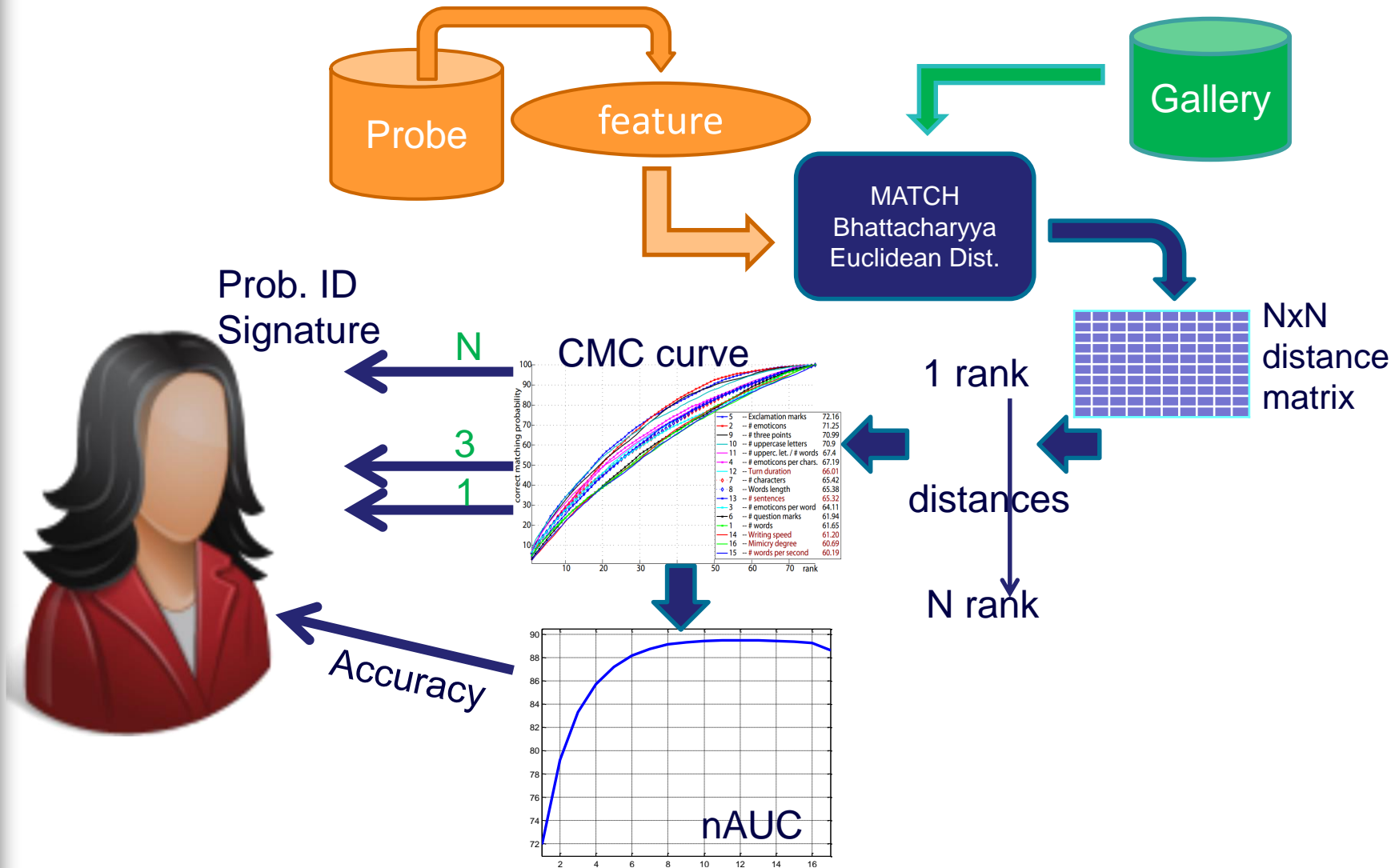
Writing speed

# Return characters

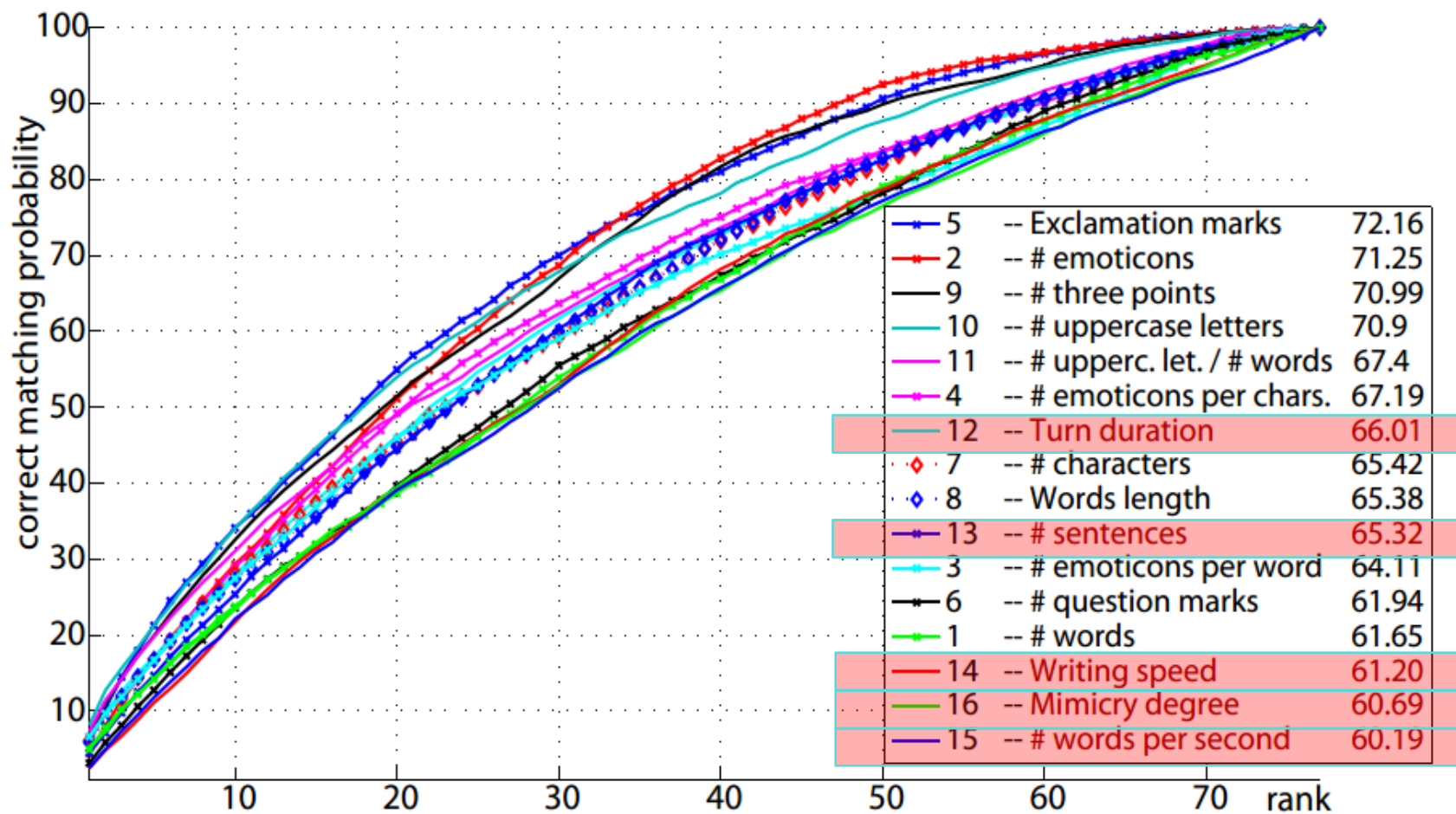
Mimicry



# Signature matching



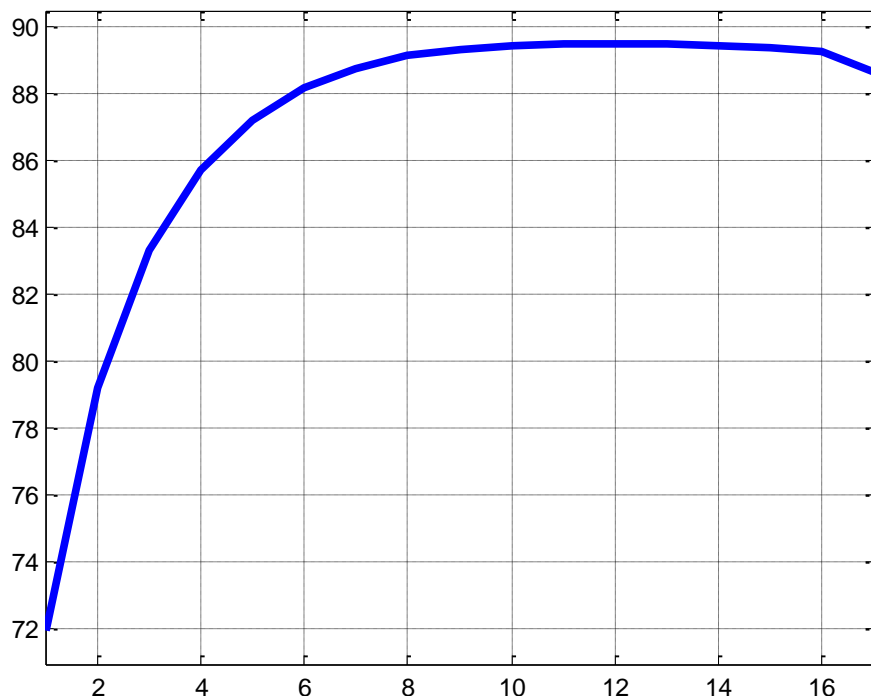
# Experiments



Single features performance

# Experiments

Feature selection:  
12 features, nAUC=90.53



normalized Area Under Curve  
(nAUC)

## Features

# exclamation marks

# emoticons

# three points

# uppercase letters

**# turn duration**

**# return chars**

average word length

**#chars per second**

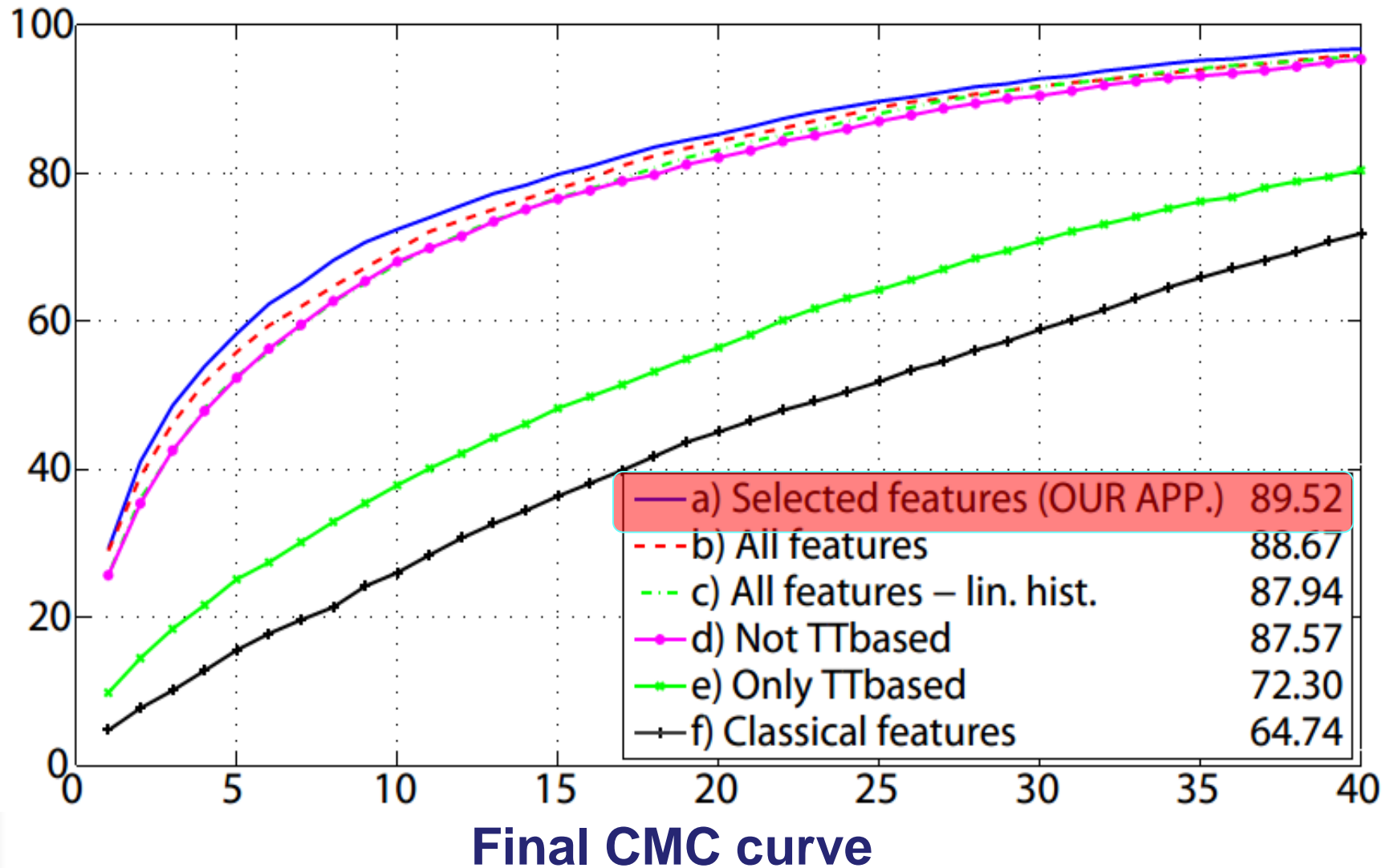
# question marks

# characters

**# words per second**

**mimicry degree**

# Experiments





# Conclusions

- Introduction of new features that account for turn-taking and mirror the features typically applied in automatic understanding of spoken conversations
- Use of turns as a basis analysis unit for the analysis of chat data and identification of their participants
- New dataset based on Skype conversations



# Future Works

- Adopt classifier to improve the results
- Compare chat and speech conversations

**"Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging" has been ACCEPTED for publication in the *Proceedings of ACM MULTIMEDIA 2012***



**Thanks for your attention!**