EVALITA 2011

Forced Alignment on Spontaneous Speech

F. Cutugno, D. Seppi, A. Origlia

Introduction

- The task consists in the alignment of a manual transcription (at words and phone levels) to a recorded speech utterance

- The automatic alignment is compared with a manual one to evaluate the accuracy in boundaries positioning

- The goal of the task is to evaluate forced alignment systems on Italian

Participants

- The SPPAS participation to Evalita 2011
  - Brigitte Bigi (CNRS - Aix-en-Provence)
- UNINA System for the EVALITA 2011 Forced Alignment Task
  - Bogdan Ludusan (University of Naples)
- SAD-based Italian Forced Alignment Strategies
  - Giulio Paci, Giacomo Sommavilla, Piero Cosi (CNR Padova)

Task Modalities

- Closed: Only provided training data could be used to train the system
  - Participants: Bigi, Ludusan, Paci/Sommavilla/Cosi

- Open: Any data could be used in training
  - Participants: Ludusan, Paci/Sommavilla/Cosi

Training Data

- 16 Italian regional varieties

- Dialogues from the CLIPS corpus (Map task and Differences test)

- 8063 training units (~ 6 hours)
  - Wav File
  - Transcription of the utterance at word level
  - Transcription of the utterance at phone level

Test Data

- Never before published dialogues recorded for the CLIPS corpus

- 89 units (10 minutes)
  - Wav File
  - Transcription of the utterance at word level

- All participants chose to present a forced alignment system integrated with their own automatic phonetic transcription step.

Evaluation

- Time mediated Alignment computed by the NIST SCLITE tool

- Word-to-word distances replaced by the following formulas:

$$D(\text{correct}) = |\, T1(\text{ref}) - T1(\text{hyp}) \,| + |\, T2(\text{ref}) - T2(\text{hyp}) \,|$$

$$D(\text{insertion}) = T2(\text{hyp}) - T1(\text{hyp})$$

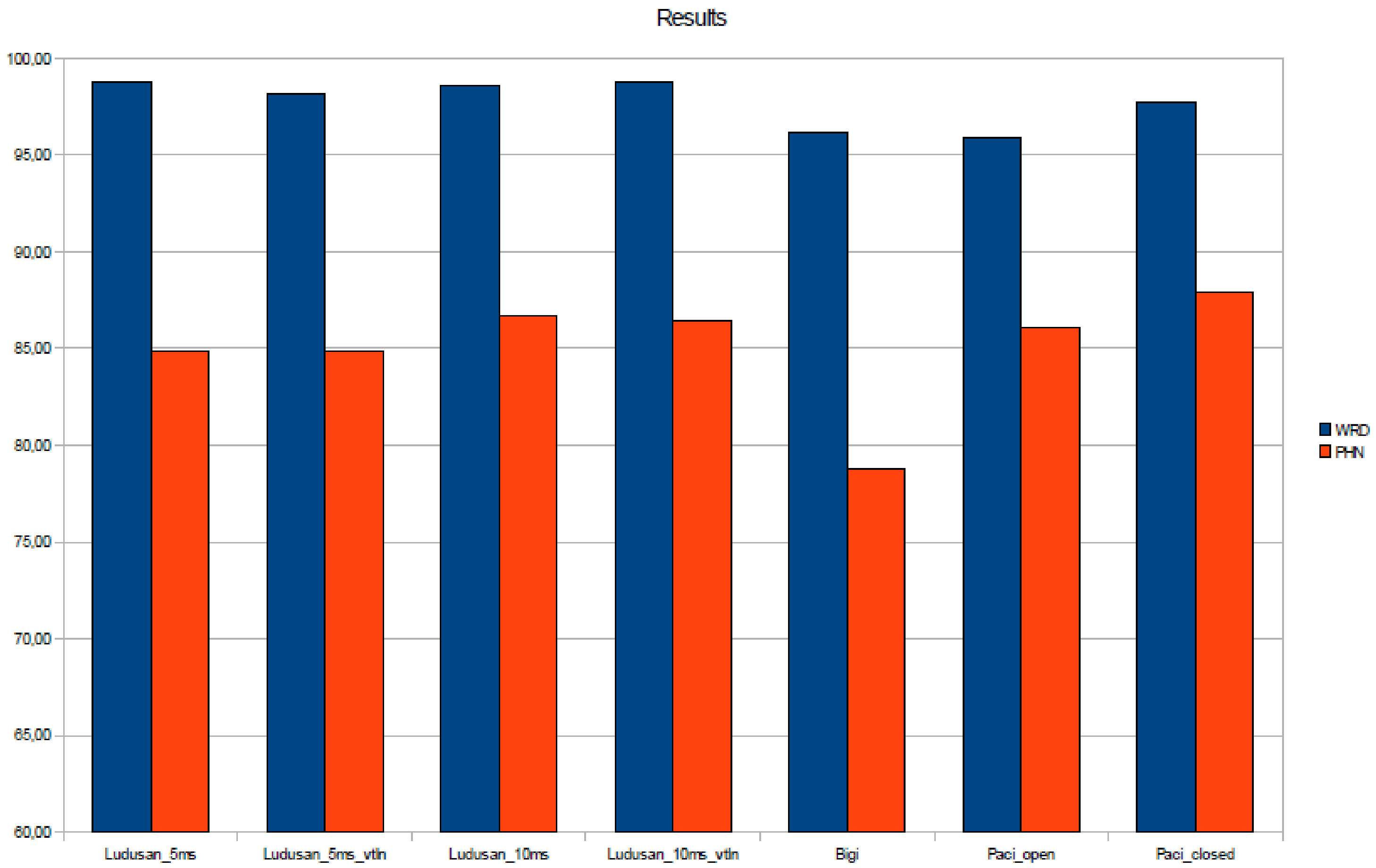
$$D(\text{deletion}) = T2(\text{ref}) - T1(\text{ref})$$

$$D(\text{substitution}) = |\, T1(\text{ref}) - T1(\text{hyp}) \,| + |\, T2(\text{ref}) - T2(\text{hyp}) \,| + 0.001$$

Phonetic transcription

- Only “clean” phones were left in the training set annotation
- Adjacent vowels were merged
- Groups of more than two vowels in the test set had alternative transcriptions (allowed by the CTM format)
- Predicted but non produced phones were not annotated in the test set
- Unpredictable phones were not annotated in the test set

Absolute results

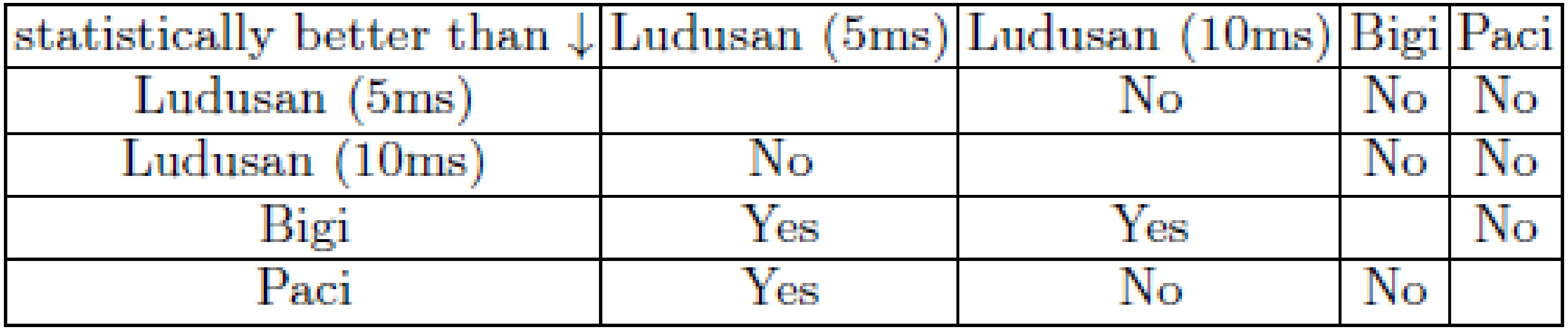


Statistical comparison

- Statistical tests performed with the NIST SC_STATS tool
  - Word alignment: Matched Pairs Sentence Segment Word Error Test
  - Phone alignment: ANOVA test

Statistical comparison

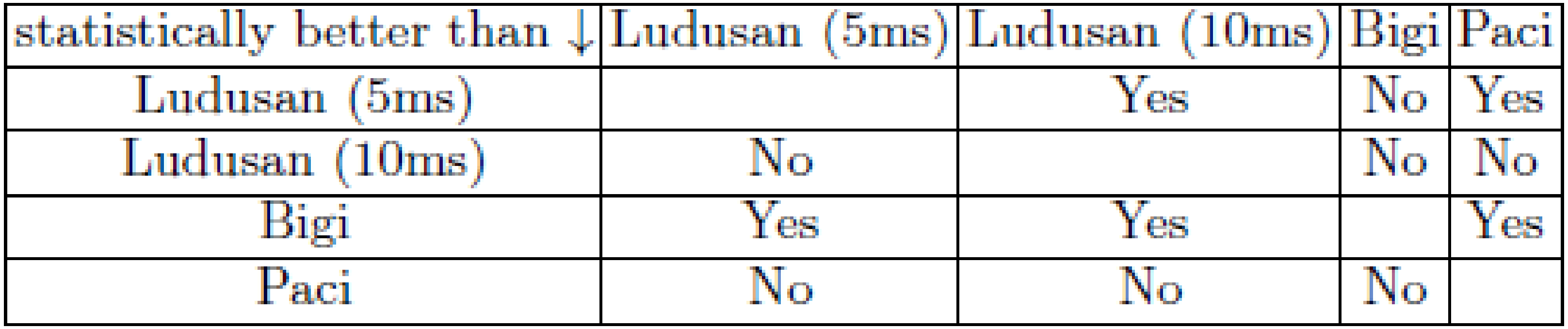
Word alignment task – closed mode
MPSS test: confidence 95%



statistically better than ↓	Ludusan (5ms)	Ludusan (10ms)	Bigi	Paci
Ludusan (5ms)		No	No	No
Ludusan (10ms)	No		No	No
Bigi	Yes	Yes		No
Paci	Yes	No	No	

Statistical comparison

Phone alignment task – closed mode
ANOVA test: confidence 95%



statistically better than ↓	Ludusan (5ms)	Ludusan (10ms)	Bigi	Paci
Ludusan (5ms)		Yes	No	Yes
Ludusan (10ms)	No		No	No
Bigi	Yes	Yes		Yes
Paci	No	No	No	

No statistically significant difference found among systems in open mode

Conclusions

- All the systems obtained very high performances even in difficult conditions

- Results are comparable to the state of the art in other languages

- Difficulties in the phone alignment task highlight the problems in annotating spontaneous speech because of reduction phenomena