Improving speech synthesis with discourse relations

THE UNIVERSITY of EDINBURGH

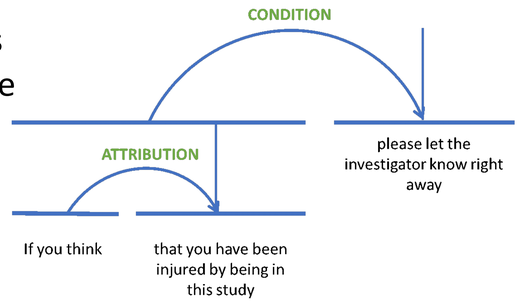
Adèle Aubin[1], Alessandra Cervone[2], Oliver Watts[1] and Simon King[1]

UNIVERSITY OF TRENTO - Italy
Department of Information Engineering and Computer Science

[1]*The Centre for Speech Technology Research, University of Edinburgh,* [2]*Department of Information Engineering and Computer Science, University of Trento*

Discourse Relations (DRs)

- **Discourse Relations** express logical **structure** of discourse
- Evidence that DRs have **acoustic** encodings



Type of DR	Definition
elaboration	S gives additional information about N *[I went to the shop $_N$][that is next to my house $_S$]*
joint	Multinuclear relation of paired Ns *[I sang $_{N1}$][and I danced $_{N2}$]*
attribution	Statement in N is reported by S *[I thought $_S$][I could do it $_N$]*
background	S gives essential information to understand N *[He ate $_N$][because he was hungry $_S$]*
contrast	Multinuclear relation where Ns are in opposition *[It seems easy, $_{N1}$][but it's not $_{N2}$]*

Hypotheses

H1 : DRs are prosodically encoded in natural speech
H2 : DRs improve naturalness of synthetic speech

We need speech data annotated with DRs

No available corpora, so we automatically annotated four audiobooks (Blizzard Challenge 2012) with the Rhetorical Structure Theory (RST)-based FastNLPParser

```
background (RightToLeft)
  TEXT:Considering the Missouri its main branch ,
  elaboration (LeftToRight)
    TEXT:it is the longest river in the world
    TEXT:-- four thousand three hundred miles .
```

Conclusions:
DRs *are* prosodically encoded and the use of DR features *does* improve synthetic speech.

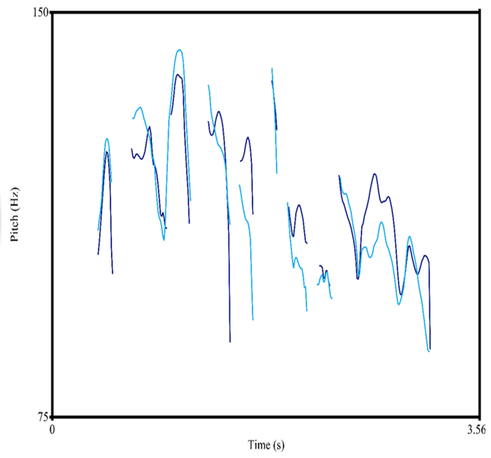
Testing H1 by acoustic analysis

- We compared utterance segments labelled with a DR with segments with no assigned DR
- Intensity in general and F0 range are significantly predicted by DRs

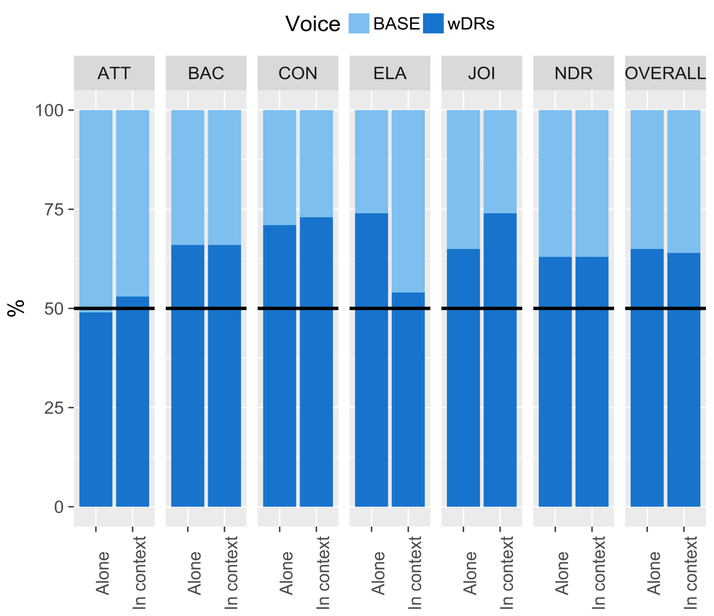
	F0			INTENSITY		
Relation	Mean	Range	SD	Mean	Range	SD
ATT		**		***	**	**
BAC	*	*		***	**	**
CON		***		***	**	**
ELA	.	***		***	**	**
JOI	***	***	**	***	**	**

Testing H2 by using Discourse Relations in speech synthesis

We used Merlin to build two neural TTS voices : BASE and wDRS
Listening test presented sentences either in isolation, or in context.



		Duration		Acoustic				
		RMSE	Corr	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	F0 corr	V/UV (%)
Valid. set	BASE	**9.027**	**0.701**	5.823	0.283	34.172	0.404	7.006
	wDRs	9.029	**0.701**	**5.713**	**0.279**	**33.824**	**0.435**	**6.958**
Test set	BASE	9.281	0.690	5.855	0.284	34.714	0.395	6.993
	wDRs	**9.265**	**0.691**	**5.744**	**0.281**	**32.204**	**0.437**	**6.931**



wDRs significantly preferred over BASE for all relation types (except ATT)