CENTRE FOR GLOBAL INTELLIGENT CONTENT

DCU at the NTCIR-11 SpokenQuery&Doc Task

David N. Racca, Gareth J.F. Jones
CNGL Centre for Global Intelligent Content
School of Computing, Dublin City University
Dublin, Ireland

Overview

CENTRE FOR GLOBAL INTELLIGENT CONTENT

— We participated in the **slide-group SQ-SCR.**

— General idea:

- Augment text-retrieval methods with prosodic features: pitch (F0), loudness, and duration.
- Compute an **acoustic score** for each term.
- Promote the rank of segments containing acoustically prominent terms.

Motivation

CENTRE FOR GLOBAL INTELLIGENT CONTENT

— Prosody:

- Rhythm, stress, intonation, duration, loudness.

— Shown useful in many speech processing tasks:

- Emotions, discourse structure, speech acts, speaker ID, topic segmentation.

— Prominent speech units stand-out from their context.

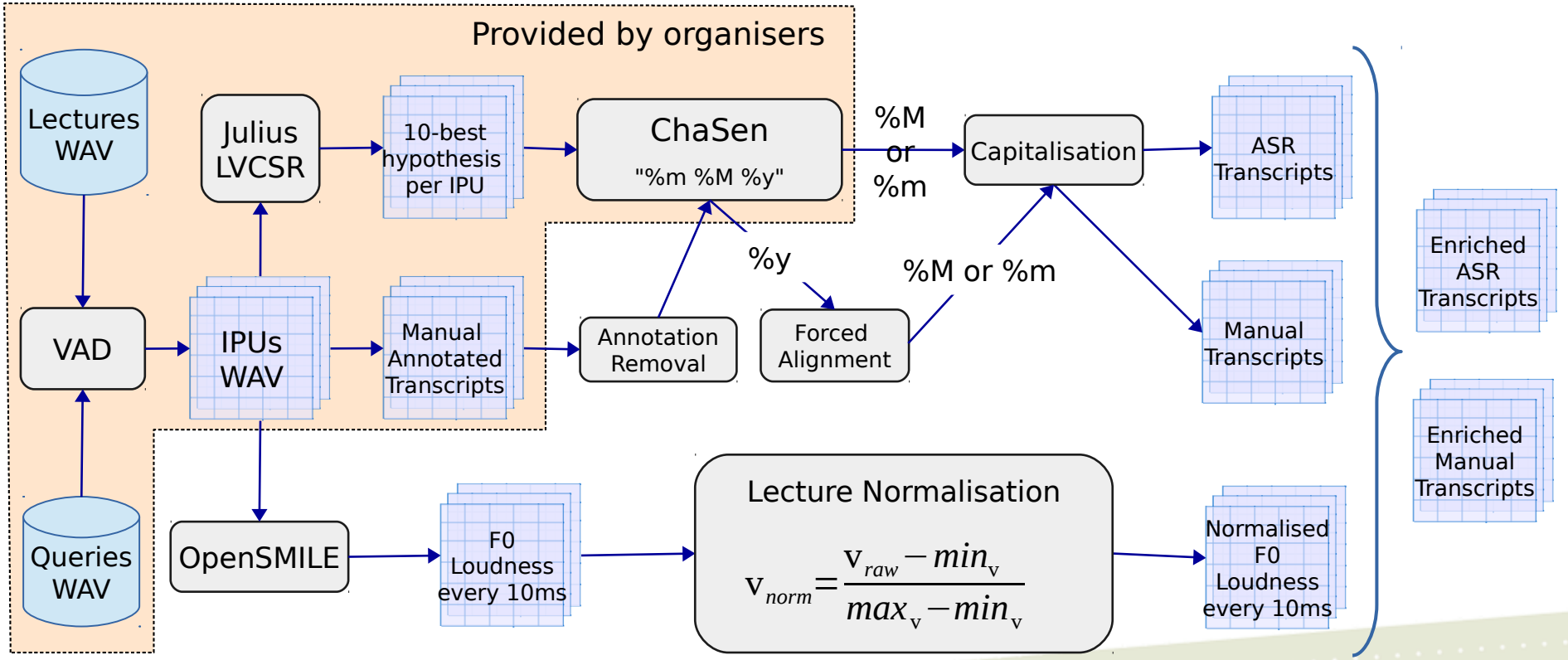
— Information status: old vs new information.

Related Work

- **Crestani [1]**: possible correlation between acoustic stress and TF-IDF scores (English).
- **Chen et al [2]**: signal amplitude and duration in a spoken document retrieval (SDR) task (Mandarin).
- **Guinaudeau [3]**: F0 and RMS energy in a topic tracking task (French).
- **Racca et al [4]**: F0, loudness, and duration in SCR (English).

Data Pre-processing

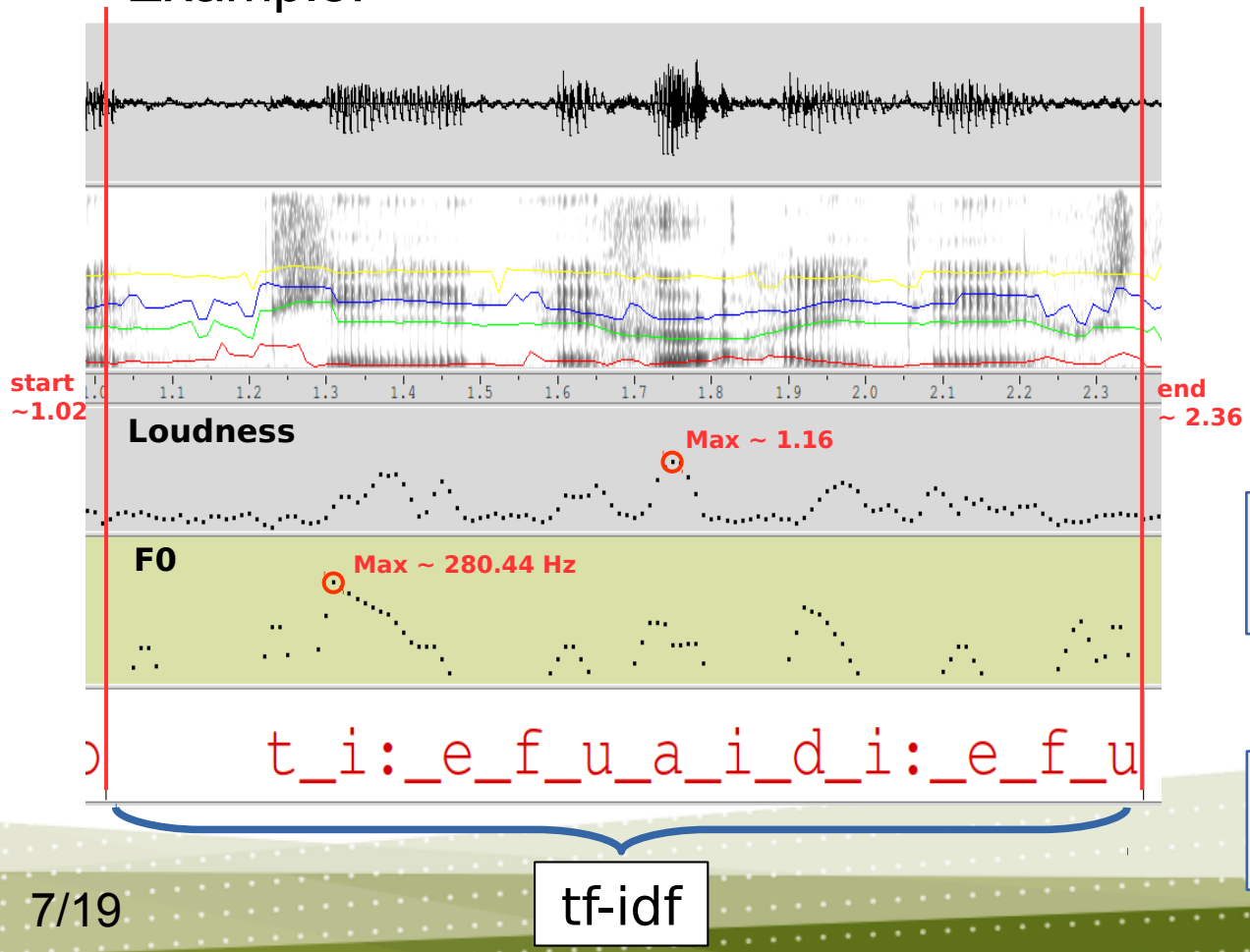
— 1-best WORD *match*, *unmatchAMLM*, and manual transcripts.



Prosodic Features

— Raw duration, lecture-normalised F0 and loudness.

— Example:



Duration

$$d = 2.36\,s - 1.02\,s = 1.34\,s$$

Lecture Normalisation

$$v_{norm} = \frac{v_{raw} - min_v}{max_v - min_v}$$

Loudness

$max(\mathbf{l}^k_{i,j}) = 1.16$ **Raw**

$max(\mathbf{l}^k_{i,j}) = 0.37$ **Normalised**

Pitch (F0)

$max(\mathbf{f0}^k_{i,j}) = 280.44\,Hz$ **Raw**

$max(\mathbf{f0}^k_{i,j}) = 0.58$ **Normalised**

Prosodic Features

—F0, loudness, and duration for the term "***i***" term in segment "***j***".

$$\mathrm{f0}(i,j) = \max_{k}\left\{ max\left(\mathbf{f0}^{k}_{i,j}\right)\right\}$$

$$\mathrm{l}(i,j) = \max_{k}\left\{ max\left(\mathbf{l}^{k}_{i,j}\right)\right\}$$

$$\mathrm{d}(i,j) = \max_{k}\left\{ \mathrm{d}^{k}_{i,j}\right\}$$

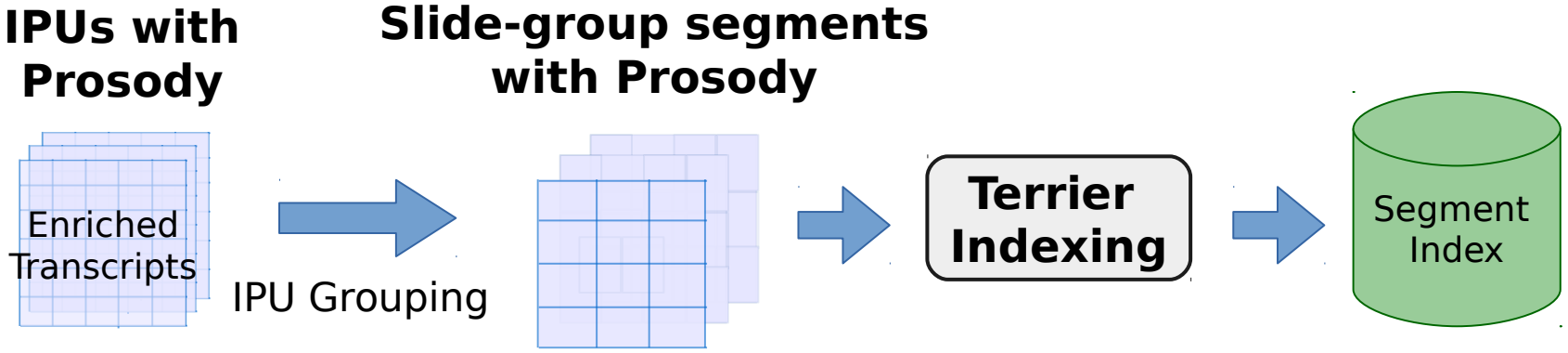
$$\mathrm{f0}_{\mathrm{range}}(i,j) = \max_{k}\left\{ max\left(\mathbf{f0}^{k}_{i,j}\right)\right\} - \min_{k}\left\{ min\left(\mathbf{f0}^{k}_{i,j}\right)\right\}$$

Acoustic Score

— We experimented with six definitions for the acoustic score of term "***i***" in segment "***j***".

$$\mathrm{ac}(i,j)=\begin{cases} \mathrm{f0}(i,j) & \textbf{Pitch [P]} \\ \mathrm{l}(i,j) & \textbf{Loudness [L]} \\ \mathrm{d}(i,j) & \textbf{Duration [Dur]} \\ \mathrm{f0}_{\mathrm{range}}(i,j) & \textbf{Pitch Range [Pr]} \\ \mathrm{l}(i,j).\mathrm{f0}(i,j) & \textbf{[LP]} \\ \mathrm{l}(i,j).\mathrm{f0}_{\mathrm{range}}(i,j) & \textbf{[LPr]} \end{cases}$$

Indexing



— Slide-group segments indexed using Terrier IR Framework.

— Index stores F0, loudness and duration for each term occurrence along with text statistics.

Retrieval

— Probabilistic model with BM25 weighting:

$$\mathrm{rel}(\boldsymbol{q}, \boldsymbol{s}_j) = \sum_{i}^{M} w(i,j)$$

— Three definitions for $w(i,j)$ were explored:

$$w(i,j) = \begin{cases} \mathrm{idf}(i,C)[\alpha . \mathrm{tf}(i,j) + (1-\alpha)\mathrm{ac}(i,j)] & \textbf{LI} \\ \dfrac{\theta_{ir} . \mathrm{tf}(i,j) . \mathrm{idf}(i,C) + \theta_{ac} . \mathrm{ac}(i,j)}{\theta_{ir} + \theta_{\mathrm{ac}}} & \textbf{G} \\ \mathrm{tf}(i,j) . \mathrm{idf}(i,C) & \textbf{TF\_IDF} \end{cases}$$

$$\mathrm{idf}(i,C) = \log\left(\frac{N}{n_i} + 1\right)$$

$$\mathrm{tf}(i,j) = \frac{k_1 . \mathrm{tf}_{i,j}}{\mathrm{tf}_{i,j} + k_1\left(1 - b + b\dfrac{\mathrm{dl}_j}{\mathrm{avdl}}\right)}$$

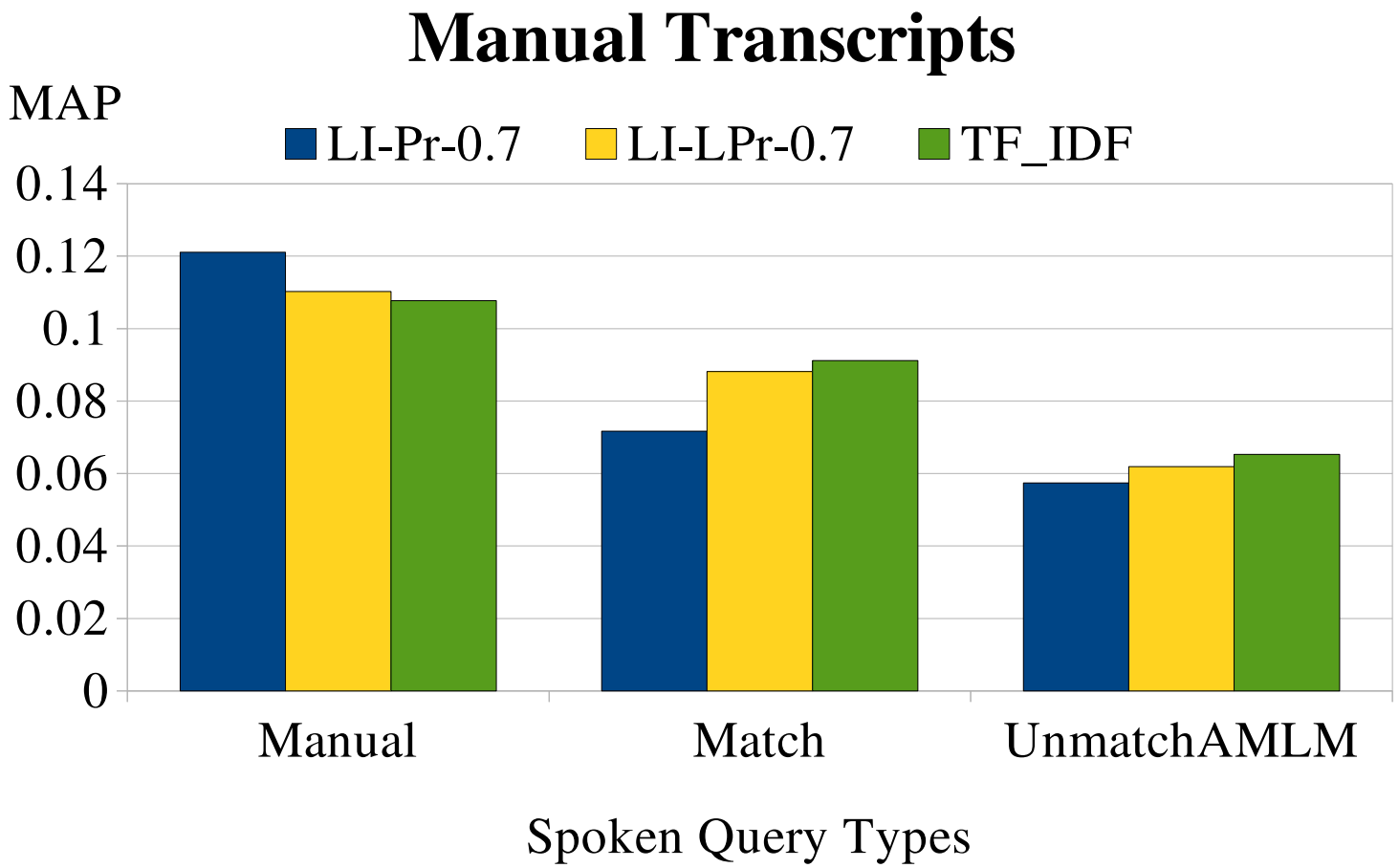
Parameter Tuning

CENTRE FOR GLOBAL INTELLIGENT CONTENT

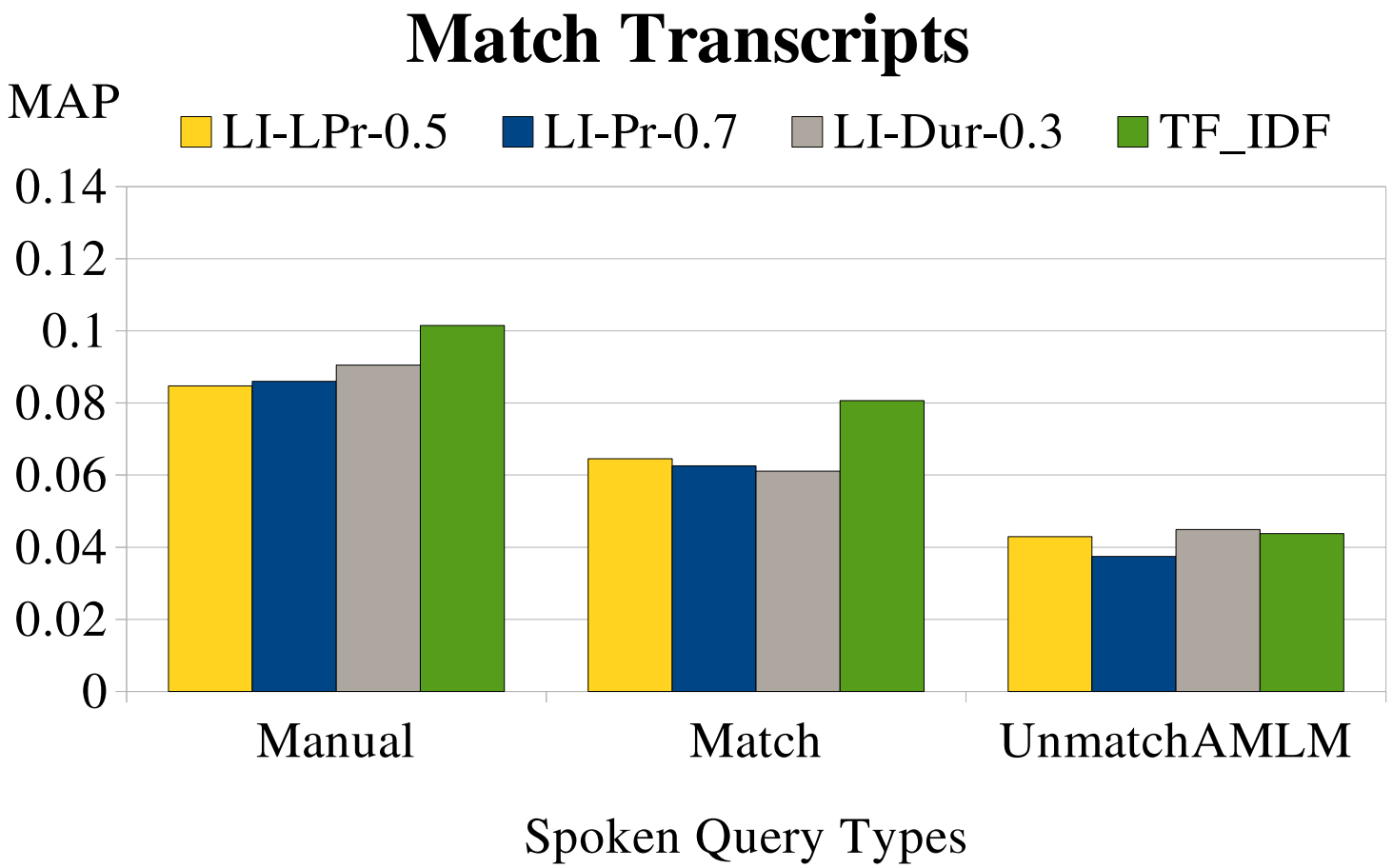
— SpokenDoc-2 passage retrieval: 120 text queries

Lecture Transcript	$w(i,j)$	$ac(i,j)$	α	θ_{ir}	θ_{ac}	uMAP	pwMAP	fMAP
Manual	LI	LPr	0.7			**.1369**	**.0976**	**.1005**
	LI	Pr	0.7			**.1369**	.0951	.0995
	G	LP		1	1	.1326	.0960	.0989
	TF-IDF					.1270	.0950	.0972
Match	LI	LPr	0.5			**.0842**	**.0508**	**.0524**
	LI	Dur	0.3			.0819	.0498	.0521
	G	Pr		1	1	.0786	.0473	.0499
	LI	Pr	0.7			.0778	.0490	.0501
	TF-IDF					.0682	.0477	.0486
UnmatchAMLM	G	P		3	1	**.0288**	.0208	.0131
	LI	LP	0.5			.0278	**.0210**	**.0135**
	LI	LPr	0.2			.0271	.0205	.0132
	LI	P	0.9			.0227	.0206	.0129
	TF-IDF					.0222	.0203	.0128

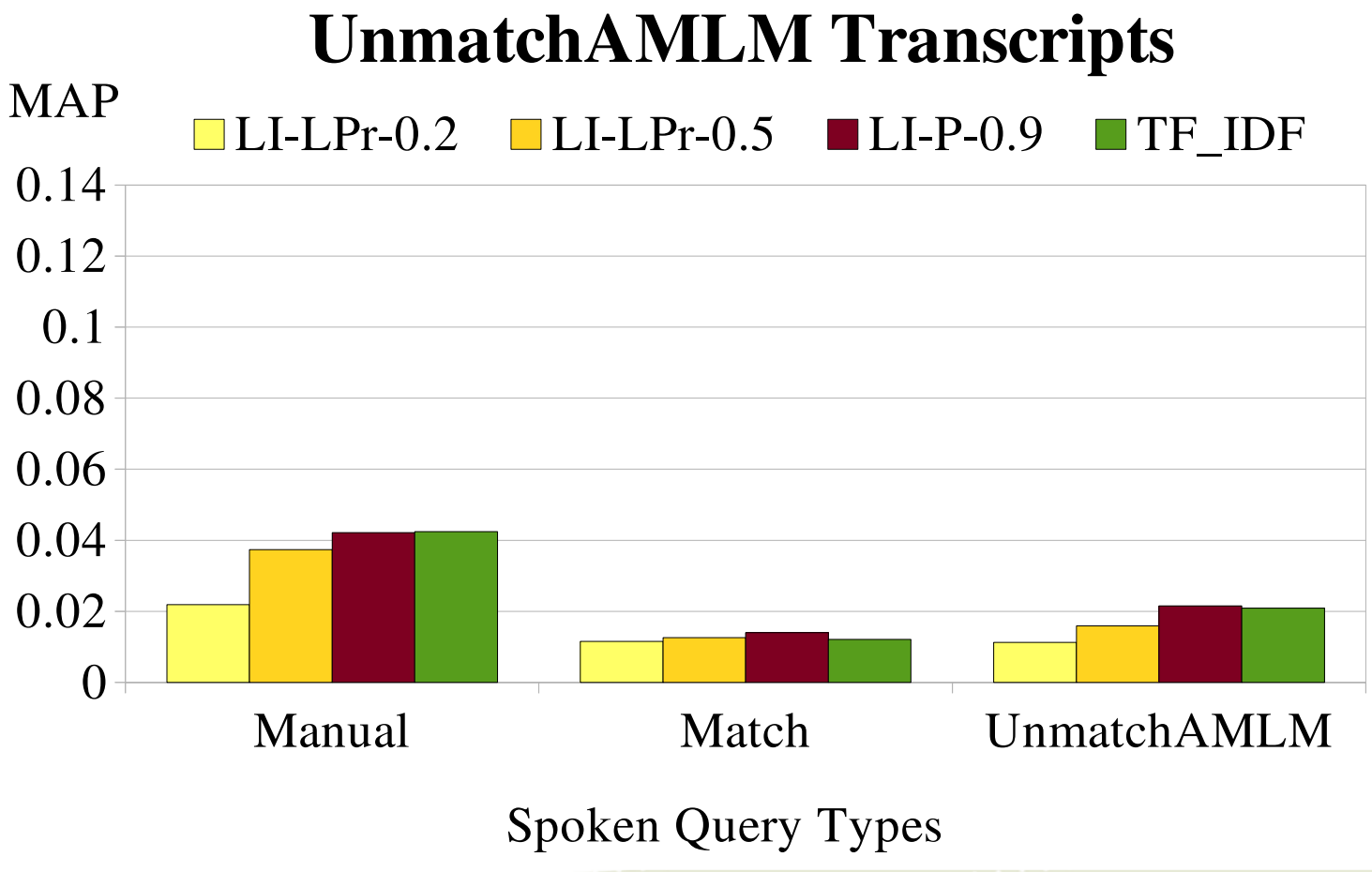
Results: SpokenQuery&Doc



Results: SpokenQuery&Doc

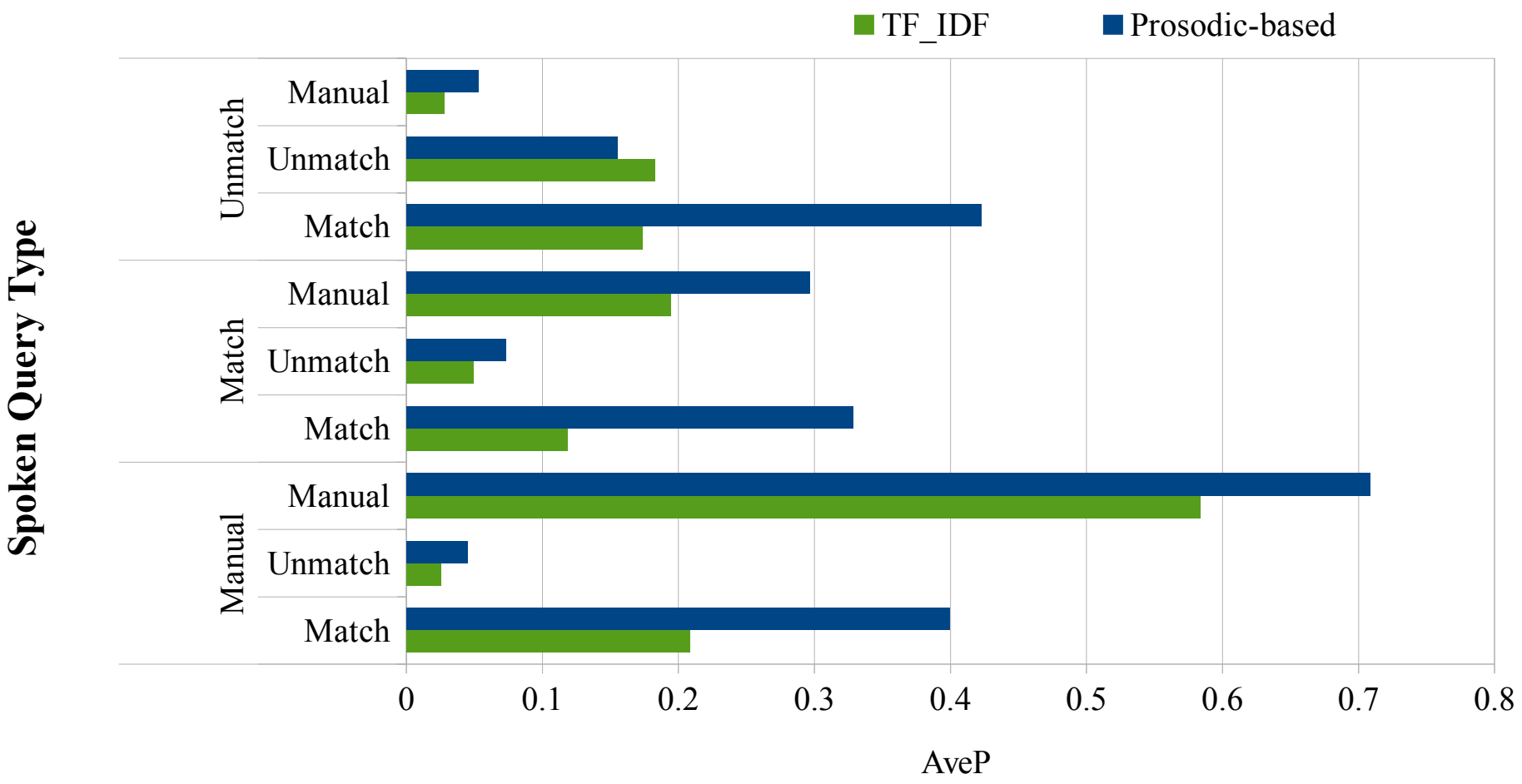


Results: SpokenQuery&Doc



Results: SpokenQuery&Doc

2 relevant segments



Conclusions & Further Work

CENTRE FOR GLOBAL INTELLIGENT CONTENT

— Continued exploring if prosodic prominence can be used to improve retrieval effectiveness.

— No significant differences between prosodic and text based runs (t student's test ~ 95% conf. level).

— Transcript quality affects retrieval effectiveness.

— Prosodic-based models may be useful for some queries/target segments:

- Future work: predict when this happens.

References

- **[1]** Crestani. Towards the use of prosodic information for spoken document retrieval. SIGIR'01, 2001.
- **[2]** Chen, et al. Improved spoken document retrieval by exploring extra acoustic and linguistic cues. INTERSPEECH'01, 2001.
- **[3]** Guinaudeau and Hirschberg. Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news. INTERSPEECH'11, 2011.
- **[4]** Racca et al. DCU search runs at MediaEval 2014 Search and Hyperlinking. MediaEval 2014 Multimedia Benchmark Workshop, 2014

Questions?