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
{dracca, gjones}@computing.dcu.ie

Introduction

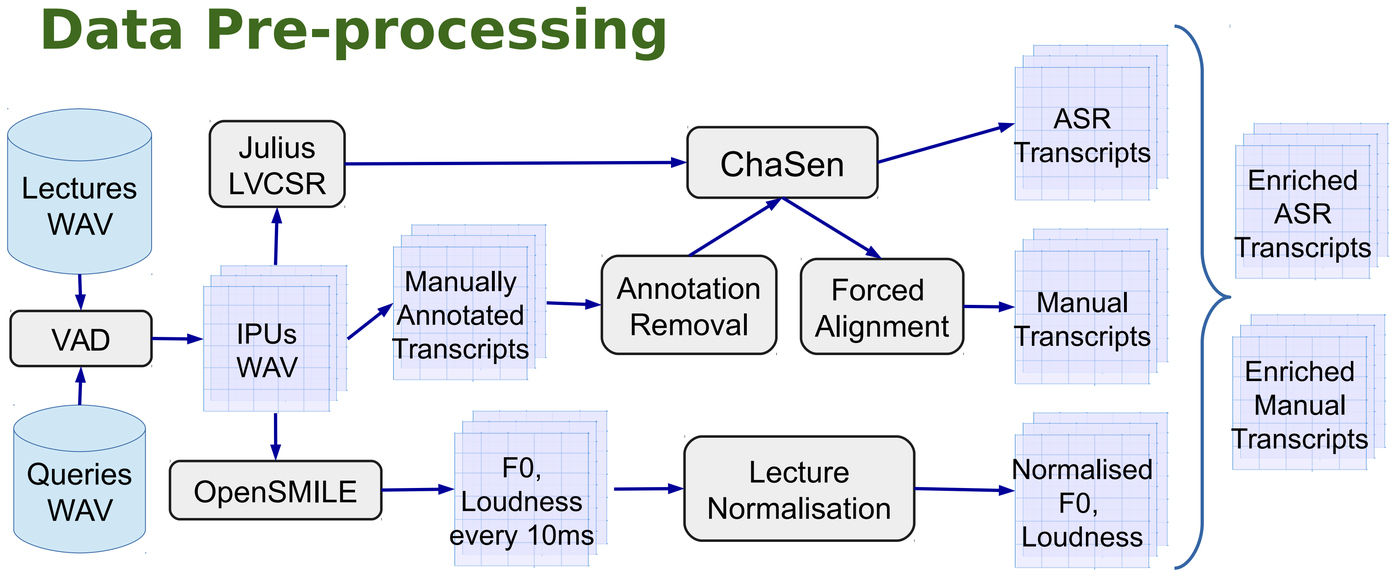
- Speech is more than a simple sequence of words.
- Prosodic variation encode rich information about:
  - emotions, discourse structure, dialogue acts, focus, emphasis, contrast, topic shifting, etc.
- We examined the potential of prosodic prominence in the NTCIR-11 SpokenQuery&Doc Task.

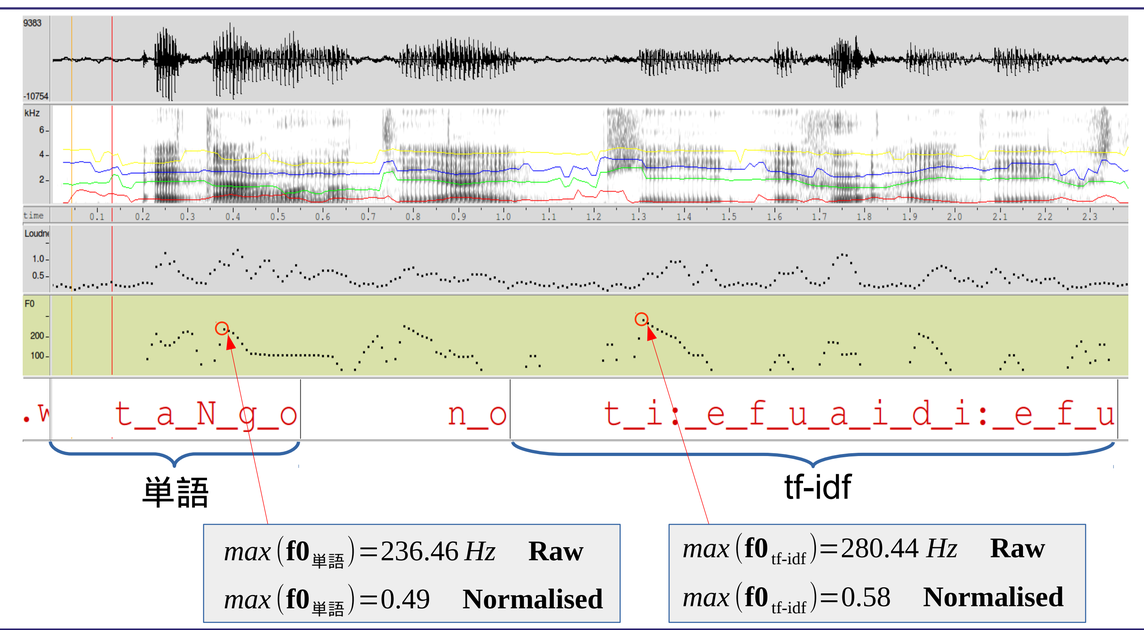
Background and Previous Work

Prosody may be useful in speech search:

- Relationship between stress and TF-IDF scores [1].
- SDR exploiting amplitude and duration [2].
- Topic tracking exploiting energy and pitch [3].
- SCR exploiting pitch, loudness, and duration [4].

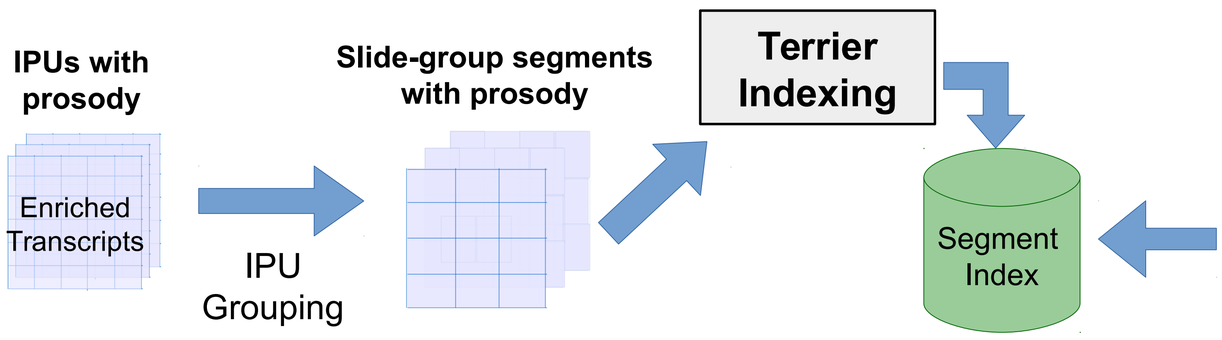
Data Pre-processing





Indexing

- Stores normalised prosodic features for each term.



Retrieval

- Increases weights of prominent terms.

Terrier Matching

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

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

$$\mathrm{tf}(i,j) = \frac{k_1 \mathrm{tf}_{i,j}}{\mathrm{tf}_{i,j} + k_1 \left(1 - b + b \dfrac{\mathrm{dl}_j}{\mathrm{avdl}}\right)} \qquad \mathrm{idf}(i,C) = \log\left(\frac{N}{n_i} + 1\right)$$

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

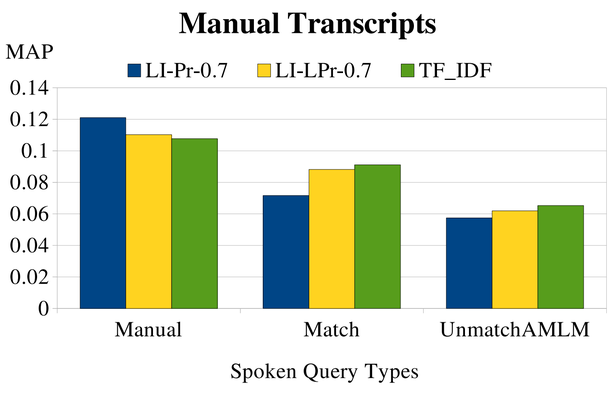
$$\mathrm{f0}(i,j) = \max_k \left\{ max(\mathbf{f0}^k_{i,j}) \right\}$$

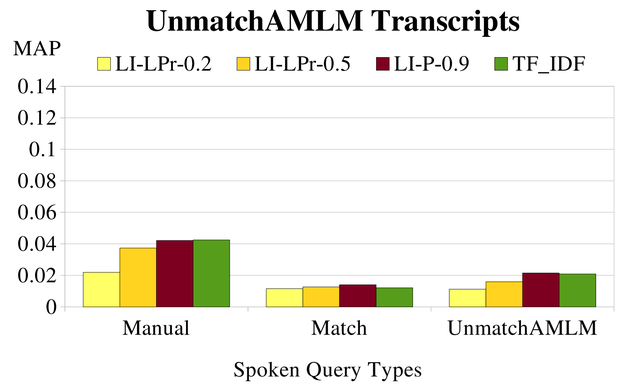
$$\mathrm{l}(i,j) = \max_k \left\{ max(\mathbf{l}^k_{i,j}) \right\}$$

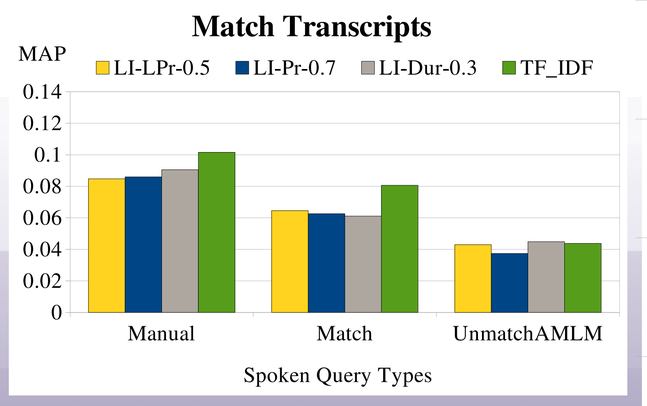
$$\mathrm{d}(i,j) = \max_k \left\{ \mathrm{d}^k_{i,j} \right\}$$

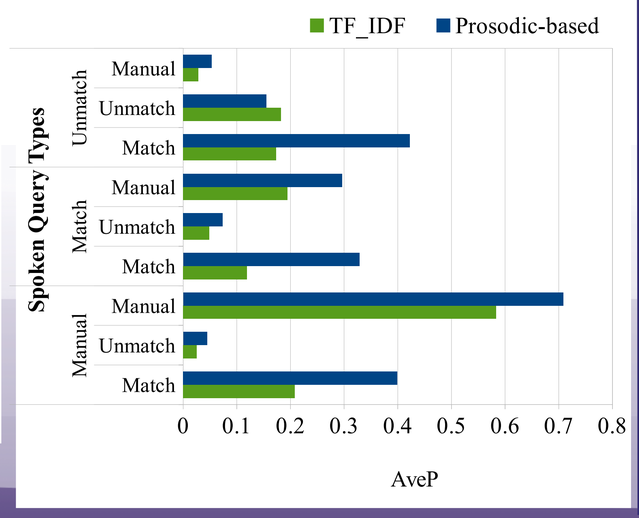
$$\mathrm{f0}_{\mathrm{range}}(i,j) = \max_k \left\{ max(\mathbf{f0}^k_{i,j}) \right\} - \min_k \left\{ min(\mathbf{f0}^k_{i,j}) \right\}$$

Results









Conclusions

- No significant differences between prosodic and text based runs.
- Transcript quality affects retrieval effectiveness.
- Prosodic-based models may be useful for certain queries.

References

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

National Development Plan 2007 - 2013

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL (www.cngl.ie)