Toward the use of information density based descriptive features in HMM based speech synthesis

Sébastien Le Maguer[1], Bernd Möbius[1], Ingmar Steiner[1,2]

[1] Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany, [2]DFKI

Introduction

- **Background**
  - Statistical TTS = huge effort assigned to acoustic modelling
  - Descriptive feature set = almost the same for each system (the one presented in [1])
- **Problem**
  - How to enrich this descriptive feature set?
- **Proposition**
  - **New descriptive feature = unpredictability of an event**
  - Based on information density & widely used in computational linguistics

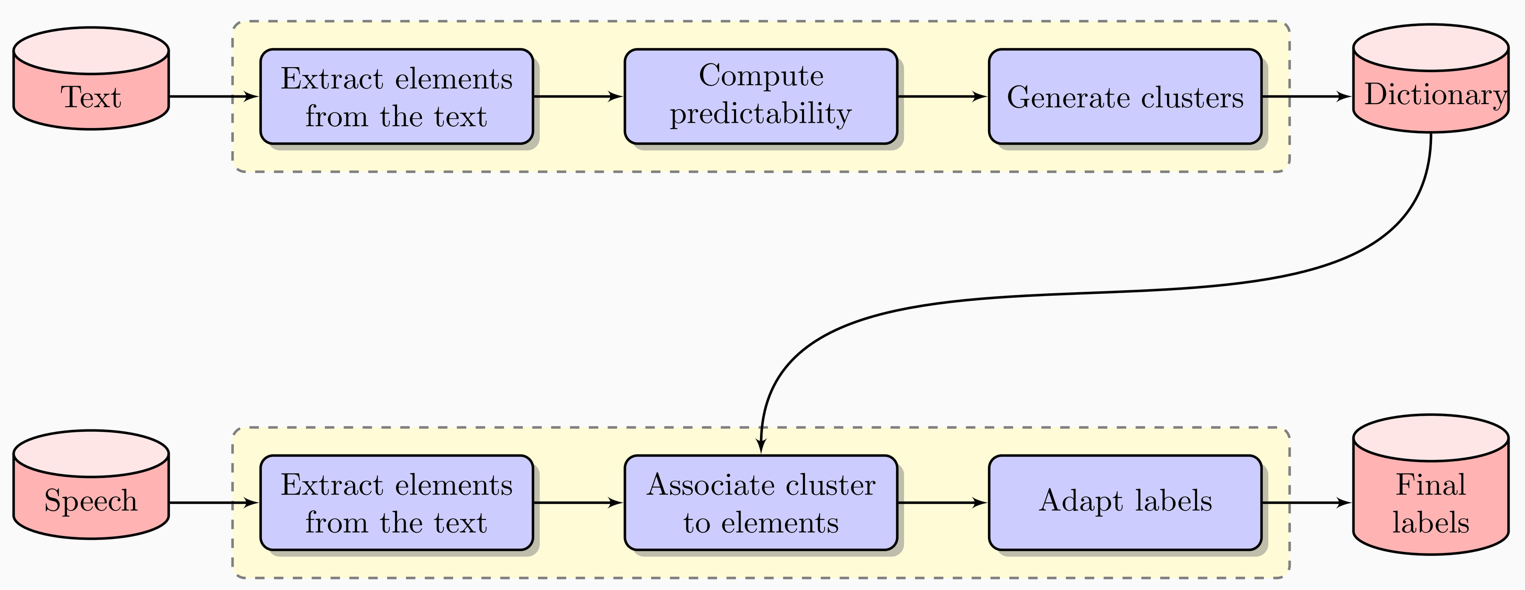
Unpredictability (Surprisal)

Once upon a ?

$$Suprisal(U_i) = -log_2(P(U_i|U_{i-1}..U_{i-1-N})) \quad (1)$$

- Predictability of a word correlates with processing effort of pronouncing this word [2]
- Same correlation found at the syllable level [3].

Process



Feature Generation

Problem: homogeneous representation (text/speech)

- **Syllable based**
  - **IPA** phoneme representation
- **Word based**
  - All **punctuation marks** are **discarded**;
  - A break mark is inserted at the **end of each paragraph**;
  - All words are converted to **lower case**

Objective Evaluation

Experimental Setup

- **Speech corpus**
  - From "Black Beauty" (2013 Blizzard Challenge),
  - 1 h (∼470 utterances) = 13 522 syl., 7038 words,
  - Segmented using EHMM + manually corrected
- **Text corpus**
  - 2013 Blizzard Challenge – "Black Beauty"
  - 82 books = 951 316 syl., 1 973 368 words
- **System setup**
  - HTS 2.2 standard configuration
  - Vocoder = STRAIGHT + MLSA filter
  - MGC (50) + LF0 (1) + BAP (25) + Δ + $\Delta\Delta$

Distance Analysis

- **6 new descriptive features**
  - syl-unpredictability (P, C, N)
  - word-unpredictability (P, C, N)
- **3 conditions**
  - *baseline*
  - *unpred_syllable*
  - *unpred_all*

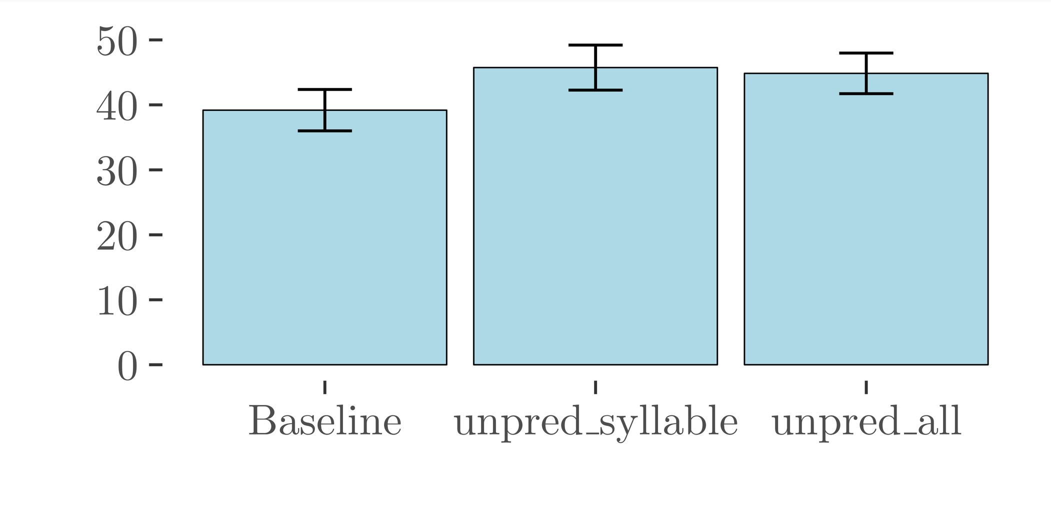
Condition	MCD	RMS-F0	VER	RMS-dur
baseline	6.45	475	15	11.1
unpred_syllable	6.33	463	14.6	10.6
unpred_all	6.33	467	14.8	10.4

Tree Analysis

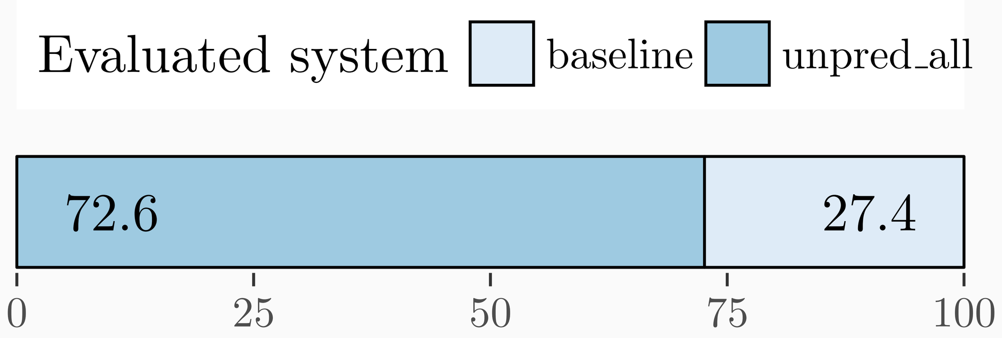
categories	baseline	unpred_syl.	unpred_all
p1	1648	1615	1594
p3	2174	2037	1175
p5	0	0	694
syl-position	**5799**	**5652**	**4056**
syl-prosody	98	128	183
syl-unpredictability	0	1657	**4163**
word-position	**7836**	**6787**	**4202**
word-prosody	2928	2817	2188
word-unpredictability	0	0	**7834**
phrase-position	1184	1573	802
phrase-prosody	**8723**	**8323**	**5892**
utterance	**7260**	**7429**	**6799**

Subjective Evaluation

MUSHRA Evaluation



Preference Evaluation



Analysis

- **AB**
  - Clear preference for the proposed system
- **MUSHRA**
  - Improvement ⇒ just a tendency
  - Evaluation: Spectrum vs. prosody?
- **Global**
  - Assumption = spectrum not impacted, prosody + natural

Conclusion

- New descriptive feature: unpredictability (widely used in computational linguistics)
- Full process to compute and apply these features
- **Objective analysis**
  - Similarity not impacted
  - Model takes this feature into account
- **Subjective evaluation**: preference for our system ⇒ which dimension?

Bibliography

[1] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings of the Speech Synthesis Workshop (SSW)*, 2002.

[2] M. Kutas, K. A. DeLong, and N. J. Smith, "A look around at what lies ahead: Prediction and predictability in language processing," in *Predictions in the Brain: Using Our Past to Generate a Future*, M. Bar, Ed. Oxford University Press, 2011, pp. 190–207.

[3] T. F. Jaeger, "Redundancy and reduction: speakers manage syntactic information density," *Cognitive Psychology*, vol. 61, pp. 23–62, 2010.