LEARNING WORD VECTOR REPRESENTATIONS
BASED ON ACOUSTIC COUNTS
MANUEL SAM RIBEIRO, OLIVER WATTS, JUNICHI YAMAGISHI
CENTRE FOR SPEECH TECHNOLOGY RESEARCH, UNIVERSITY OF EDINBURGH, UK
m.f.s.ribeiro@sms.ed.ac.uk

INTRODUCTION

Good representations of suprasegmental units (e.g., words, syllables) are essential for the natural generation of speech prosody in text-to-speech systems.

- But most systems typically use shallow representations of context (e.g., positional features or POS tags).

This paper proposes a method to learn acoustically-motivated vector representations of words and syllables.

- This is a data-driven method that learns a vector space model (VSM) over a parallel corpus of text and speech.
- VSMs are based on the distributional hypothesis: words that occur in similar contexts have similar meaning.

VECTOR REPRESENTATIONS

- Define a vocabulary set \( V \), an acoustic class set \( A \), and a window of size \( w \).
- Build co-occurrence matrix \( M \in \mathbb{R}^{|V| \times |A|} \).
- For each row, normalize each sub-vector s.t. we have \( w \) probability distributions.
- Find the SVD of \( M \), s.t. \( M = UVV^T \) and take \( k \) left singular values.
- We are left with a matrix \( \hat{U} \in \mathbb{R}^{|V| \times k} \).

Integration of multiple sources of information:

- \( f_0 \) and energy cluster-based approach – 8 DCT coefficients, 20 clusters/classes (+1)
- \( f_0 \) mean-based approach – range: [100-300], bin: 2Hz, classes: 100 (+3)
- energy mean-based approach – range: [3-7], bin: 0.05, classes: 80

DISCRETIZATION APPROACHES

- \( \hat{U} \) is the resulting matrix.

Figure: Co-occurrence matrix.

Experiments vary three main factors regarding the learned representations:

- Discretization method: cluster, mean.
- Acoustic signal: \( f_0 \), energy.
- Linguistic level: words, syllables.

Integration of multiple sources of information:

![Integration of multiple sources of information](image)

We have used expressive audiobook data. Approximately 18 hours over 13000 utterances, with roughly 220k word and 300k syllable tokens.

Word representations use a vocabulary of 4468 word types with 8.7% mapped to UNK. Syllables use a vocabulary of 3447 types with 1.9% of total tokens mapped to UNK.

In general, we find that:

- cluster < mean < cluster+mean
- energy < \( f_0 \) < \( f_0 \)-energy
- syllable < word < word+syllable

FUTURE WORK

- Additional discretization methods (e.g., SLAM), linguistic levels (e.g., phrases), or acoustic signals (e.g., jitter, shimmer).
- Evaluate speaker dependency of vector representations.
- Effect of database size on vector representations.
- Evaluation of vectors as replacement of knowledge-based features in low-resource languages.

CONCLUSION

Results indicate that the proposed data-driven vector representations improve the performance of a TTS system.

In general, we have observed that the more discretization approaches, acoustic signals, and levels of linguistic analysis are incorporated into a TTS system via these count-based representations, the better that TTS system performs.