Unsupervised Visualization of Under-resourced Speech Prosody

Moses Ekpenyong , Udoinyang Inyang , Emem Obong Udoh

PII: S0167-6393(17)30225-X
DOI: 10.1016/j.specom.2018.04.011
Reference: SPECOM 2562

To appear in: *Speech Communication*

Received date: 17 June 2017
Revised date: 10 January 2018
Accepted date: 30 April 2018

Please cite this article as: Moses Ekpenyong , Udoinyang Inyang , Emem Obong Udoh , Unsupervised Visualization of Under-resourced Speech Prosody, *Speech Communication* (2018), doi: 10.1016/j.specom.2018.04.011

Unsupervised Visualization of Under-resourced Speech Prosody

Moses Ekpenyong [a,*], Udoinyang Inyang[a], EmemObong Udoh[b]
[a]*Department of Computer Science, University of Uyo, P.M.B. 1017, Uyo 520003, Akwa Ibom State, Nigeria*
[b]*Department of Linguistics and Nigerian Languages, University of Uyo, P.M.B. 1017, Uyo 520003, Akwa Ibom State, Nigeria*

[*]Corresponding author.
E-mail address: mosesekpenyong@{uniuyo.edu.ng, gmail.com} (M. Ekpenyong).

Abstract
In this paper, an unsupervised visualization framework for analyzing under-resourced speech prosody is proposed. An experiment was carried out for Ibibio – a Lower Cross Language of the New Benue Congo family, spoken in the Southeast coastal region of Nigeria, West Africa. The proposed methodology adopts machine learning, with semi-automated procedure for extracting prosodic features from a translated prosodically stable corpus 'The Tiger and the Mouse' – a text corpus that demonstrates the prosody of read-aloud English. A self-organizing map (SOM) was used to learn the classification of certain input vectors (speech duration, fundamental frequency: F0, phoneme pattern (vowels only), tone pattern), and provide visualization of the clusters structure. Results obtained from the experiment showed that duration and F0 features realized from speech syllables are indispensable for modeling phoneme and tone patterns, but the tone input classes revealed clusters with well separated boundaries and well distributed component planes, compared to the phoneme input classes. Further, except for very few outliers, the map weights were well distributed with proper neighboring neuron connections across the input space. A possible future work to advance this research is the development of the language's corpus, for the discovery of prosodic patterns in expressive speech.

Keywords: Machine learning; pattern analysis; self-organizing map; speech prosody; tone modeling.

1. Introduction

Classically, prosody is concerned with speech features (within a larger domain) and spans more than one segment. It manifests in syllables rather than individual phonetic units such as vowels and consonants, and is often referred to as supra-segmental [1]. These features are useful, as they offer clues to (i) linguistic functions such as intonation, tone, stress, and rhythm; (ii) revealing the speaker's gender, identity and linguistic background; (iii) revealing the speaker's emotional state (e.g., anger, sadness, joy); (iv) understanding the form of an utterance – the presence of irony or sarcasm, emphasis, contrast, and focus, or other aspects of language that may not be encoded by grammar or by choice of vocabulary. The study of supra-segmental requires a robust visualization of intonation, stress, rhythm, and syllable structure [2]. Intonation is the pitch movement over vowels and semi-vowels only, but traditional algorithms consider pitch on all voiced segments indiscriminately, thus confusing the visualization process.

In the study of speech prosody, it is usual to distinguish between auditory measures (subjective impressions perceived or produced in listeners mind) and acoustic measures (physical sound properties that may be objectively or mechanically measured). These measures do not linearly correlate [3] – because our ears respond differently to low and high frequencies in terms of how far apart they are, and changes in low frequency range are very obvious but not so obvious in high frequency range. Most studies of prosody however rely on auditory analysis using auditory scales. Prosodic features correspond to the