Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

Crowd-sourcing prosodic annotation☆

Q1 Jennifer Cole[a,b,*], Timothy Mahrt[c], Joseph Roy[a]

[a] University of Illinois at Urbana-Champaign, Department of Linguistics 4080 Foreign Language Building, 707 S Mathews Avenue, MC-168, Urbana, Illinois 61801, USA
[b] Northwestern University, Department of Linguistics, 2016 Sheridan Road Evanston, Illinois 60208, USA
[c] Laboratoire Parole et Langage, UMR 7309 CNRS, Aix-Marseille Université, 5 avenue Pasteur BP 80975, Aix-en-Provence 13604, France

Received 30 August 2016; received in revised form 13 February 2017; accepted 14 February 2017

Abstract

Much of what is known about prosody is based on native speaker intuitions of idealized speech, or on prosodic annotations from trained annotators whose auditory impressions are augmented by visual evidence from speech waveforms, spectrograms and pitch tracks. Expanding the prosodic data currently available to cover more languages, and to cover a broader range of unscripted speech styles, is prohibitive due to the time, money and human expertise needed for prosodic annotation. We describe an alternative approach to prosodic data collection, with coarse-grained annotations from a cohort of untrained annotators performing rapid prosody transcription (RPT) using LMEDS, an open-source software tool we developed to enable large-scale, crowd-sourced data collection with RPT. Results from three RPT experiments are reported. The reliability of RPT is analysed comparing kappa statistics for lab-based and crowd-sourced annotations for American English, comparing annotators from the same (US) versus different (Indian) dialect groups, and comparing each RPT annotator with a ToBI annotation. Results show better reliability for same-dialect annotators (US), and the best overall reliability from crowd-sourced US annotators, though lab-based annotations are the most similar to ToBI annotations. A generalized additive mixed model is used to test differences among annotator groups in the factors that predict prosodic annotation. Results show that a common set of acoustic and contextual factors predict prosodic labels for all annotator groups, with only small differences among the RPT groups, but with larger effects on prosodic marking for ToBI annotators. The findings suggest methods for optimizing the efficiency of RPT annotations. Overall, crowd-sourced prosodic annotation is shown to be efficient, and to rely on established cues to prosody, supporting its use for prosody research across languages, dialects, speaker populations, and speech genres.

Keywords: Prosody; Annotation; Crowd-sourcing; Generalized mixed effects model; Inter-rater reliability; Speech transcription

1. Introduction

Investigations into the prosody of spoken languages—whether done in the service of describing languages, theorizing about language structure, or modelling spoken language processing—rely on the analysis of prosodic data, which comes very often in the form of prosodic annotation. Important early discoveries about prosody, such as the role of phrasal prominence in marking information structure in English (Bolinger, 1954; Halliday, 1967; Chafe,

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.
* Corresponding author at: Northwestern University, Department of Linguistics, 2016 Sheridan Road Evanston, Illinois 60208, USA.
E-mail address: jennifer.cole1@northwestern.edu (J. Cole), timmahrt@gmail.com (T. Mahrt), jroy042@illinois.edu (J. Roy).

http://dx.doi.org/10.1016/j.csl.2017.02.008
0885-2308/

1987), drew on data in the form of native speaker intuitions of idealized speech. Other work, including most current research, examines recordings of elicited or spontaneous speech for which prosodic annotations are produced by trained annotators based on auditory impression alone (e.g., Crystal, 1969; Bolinger, 1982), or augmented by visual evidence from pitch tracks, waveforms, and spectrogram displays (e.g., Bolinger, 1958; Ladd, 1980; Pierrehumbert, 1980; Gussenhoven, 1984; Wightman et al., 1992; Grabe and Post, 2002; Calhoun et al., 2010).

Prosodic annotations of recorded speech have many obvious advantages over the more purely subjective and impressionistic annotations of earlier work. For instance, the recordings can be submitted to multiple independent annotators, with inter-annotator agreement rates offering a measure of the reliability of the annotation (Pitrelli et al., 1994; Yoon et al., 2004; Breen et al., 2012). In addition, and most obviously, the presence of an audio recording means that acoustic correlates of the prosodic features labelled by annotators can be measured and identified. The distribution of these acoustic cues can then be examined to assess the contrastive status of the annotated prosodic features, and to identify systematic patterns of contextual variation in the phonetic expression of those features. Furthermore, prosodically annotated recorded utterances are also useful as stimuli for research on the perception of prosodic features and their influence on sentence and discourse comprehension.

Unfortunately, the advantages of working with prosodic annotations of recorded speech are available only for languages and dialects for which there exists a prosodic annotation standard, an available supply of trained annotators, and the necessary resources of money and time to perform the annotation and its validation by means of reliability analysis. In practice, these requirements have restricted prosody research primarily to the "big" languages of the world, i.e., those that have the support of a large community of researchers with access to research funding, and to the standard varieties for which annotation systems have been developed. Thus, in comparison to the growing body of prosody research on e.g., standard and regional varieties of Dutch, English, French, German, Italian, Japanese, Portuguese, and Spanish, there remains scant research on the vast majority of languages, notably, for most of the smaller, "under-resourced" languages and for non-standard and L2 varieties, but also for some languages with large speaker populations and a body of linguistic scholarship, such as Arabic and Russian.[1]

Here we present *rapid prosody transcription* (RPT) as an alternative methodology for prosodic annotation, one that sidesteps the limitations of traditional annotation methods by using untrained annotators in place of trained experts, and coarse-grained prosodic features in place of a larger and more detailed feature inventory (Mo et al., 2008). The simplicity of an RPT annotation, deriving from its use of only two binary features, one for prominence and one for prosodic phrase boundaries, is offset by more nuanced distinctions that are revealed when annotations of the same speech materials are aggregated over a group of annotators. Differences among annotators in their rating of words as prominent or as preceding a prosodic boundary reveal complex patterns of association between prosodic features and the cues to these features that are present in the speech signal and in the broader linguistic context of the utterance (Cole et al., 2010a, 2010b)

As described in more detail below, RPT requires no training or special knowledge of prosodic theory, and RPT annotation tasks can be performed without supervision. RPT is not the first annotation method to rely on "näve" annotators; similar methods have been used to obtain prosodic ratings/judgements for a variety of research interests in prior work (de Pijper and Sanderman, 1994; Swerts, 1997; Streefkerk et al., 1997, 1998; Buhmann et al., 2002; Wagner, 2005). The use of untrained annotators confers an advantage in that RPT can be performed outside the research laboratory, with speech materials presented via audio files that are accessed online, and annotations entered digitally and relayed to the researcher through the internet. Consequently, RPT annotators can be recruited from any location with internet access. These properties of RPT enable its use with any language variety and any genre of speech for which an orthographic transcript can be produced. Annotators can be recruited online through crowd-sourcing platforms or other internet resources, allowing researchers to investigate prosody through the lens of annotation data from a much larger sample of the language community. (Hasegawa-Johnson et al., 2015).

This paper reports on three large-scale RPT studies, one using RPT in a lab setting, and two using RPT deployed over the internet with annotators recruited through a crowd-sourcing platform. Our goal in this paper is to evaluate

[1] This is not to imply that there are no prosodic analyses of languages and varieties outside the privileged group that includes standard, L1 English. Prosodic analyses of other languages and varieties are few but are also gaining in number in the literature, critically, as supported by the development of prosodic annotation standards for those languages (Gussenhoven, 2004; Jun, 2006, 2014) and by guidelines for prosodic field work (Jun and Fletcher, 2014; Arvaniti, 2016).

Please cite this article as: J. Cole et al., Crowd-sourcing prosodic annotation, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.02.008