A CUE-BASED APPROACH TO PROSODIC DISFLUENCY ANNOTATION

Alejna Brugos[1,4], Alison Langston[2], Stefanie Shattuck-Hufnagel[3], Nanette Veilleux[4]

[1] Boston University, [2] Wellesley College, [3] MIT, [4] Simmons College
abrugos@bu.edu, alangsto@wellesley.edu, sshuf@mit.edu, veilleux@simmons.edu

ABSTRACT

This paper elaborates a proposal for labelling prosodic disfluencies in American English, in conjunction with the ToBI framework for prosodic labelling. Incorporating disfluency annotation ideas developed for other languages, and for stuttered speech, the proposal introduces explicit disfluency-related labels into the Break Index tier, providing a more fine-grained categorization of the prosodic disfluency type than established ToBI disfluency labels. In addition, it explicitly labels 'speech errors' (e.g. word and sound sequencing errors) even when not disfluent, enabling separate study of these phenomena, and of their interaction. The paper further explores data labelled using this system by two independent labellers (recordings of spontaneous speech by 5 female speakers of Mainstream American English). The relative frequency of specific disfluency types, alone or in combination with others, reflects a high degree of diversity. Differences in disfluency patterns used by individual speakers are discussed.

Keywords: ToBI, disfluencies, prosody, prosodic annotation, cues

1. INTRODUCTION

Disfluencies, i.e. perceived disruptions to the smooth flow of speech, are frequent occurrences in naturally produced spoken language, particularly in spontaneous speech. The study of disfluencies and non-disfluent speech errors has the potential to provide insight into speech planning processes. Prosodic disfluencies may arise from several different causes, including detected speech errors or planning difficulties, or potentially from stylistic decisions. Not all disfluencies are the same, and different types of disfluencies have been shown to affect speech perception and recall, and to do so in different ways. Filled pauses have been shown to aid both processing [17] and recall [18], and likewise silent pauses [21]; however, repetitions have not necessarily shown this effect [20]. Disfluent interruptions/distortions of prosodic constituents are distinct from non-disfluent morphosyntactic and segmental 'speech errors', such as substitutions and exchanges, and show a differing distribution in discourse [34], although the two can also occur together. For these reasons, when annotating the prosody of naturally produced speech, it is useful not only to flag the occurrence of perceived disfluencies, but also to annotate some details of the specific nature of each disfluency occurrence. The ToBI system [6, 32] has in place conventions for labelling disfluencies, but labellers and researchers have found the constraints on those labels to be at times both challenging and inadequate for capturing the variability. This paper proposes a more user-friendly system for transcribing disfluencies than the conventions for MAE ToBI currently offer.

2. BACKGROUND

ToBI (for Tones and Break Indices) [6, 7, 8, 32] is a system for the annotation of spoken prosody. ToBI is based largely on the Tone category work of Pierrehumbert [26] and collaboration with Beckman [9], and on the "break indices" work of Price and colleagues [27]. ToBI uses time-aligned text-based annotation of a speech file, typically in tiers. Conventions [7, 8] call for the following 4 tiers[1]:

1) **words:** the orthographic tier for lexical words
2) **tones:** for categorical labels for tonal markings of prominence (pitch accents) and boundary markings (edge tones)
3) **breaks:** for degree of perceived disjuncture between 0 (lowest) and 4 (highest).
4) **misc** (miscellaneous): for additional speech features and comments

Disfluency labelling in established ToBI conventions, based largely on a proposal by Nakatani and Shriberg [25], is distributed across all 4 tiers, but primarily in the breaks tier, with the use of the **p diacritic** [7: p. 36]:

1p: an abrupt cutoff before [a] repair, or as if stopping to permit a repair or restart of some kind
2p: a hesitation pause or prolongation of segmental material where there is no phrase accent perceived in the intonation contour
3p: a hesitation pause or a pause-like prolongation where there is a phrase accent in the tone tier.

ToBI conventions call for one disfluency label in the tones tier: %r, for a restarted prosodic phrase after a previous phrase was not finished with typical

boundary cues. Regions of disfluency, or other disfluent events not captured using the p diacritic, such as repairs or restarts, can be marked in the miscellaneous tier, but labels for these are not well established. Filled pauses (typically "um" or "uh" in Mainstream American English) are at times flagged with p diacritics, but often are labelled only with break indices corresponding to fluently produced speech, and therefore only indicated as a filled pause via the words tier label.

While functional, existing ToBI conventions for disfluencies have several disadvantages: 1) they obligatorily relate p-diacritic marked disfluencies to well-formed prosodic structure (1p, 2p, 3p), yet disfluencies are often ambiguous with respect to the intended well-formed target, 2) they don't adequately distinguish different specific disfluency cues: E.g. pause, prolongation, filled pauses, etc. and 3) such cues and phenomena appear in a wide variety of combinations, not necessarily in a fixed relationship.

All of these result in challenges for labellers, and as a result, labellers often fail to use disfluency markers consistently, or skip labelling disfluencies altogether. To address these challenges, we propose modifications to disfluency labelling conventions in ToBI. A substantial body of research into the production and perception of disfluencies, in many languages and language varieties, provides a rich field of potential annotation systems to consider. The following proposal makes reference to several of these, and we expect future refinements of the system to incorporate aspects of additional systems.

3. CURRENT PROPOSAL: NEW TOBI DISFLUENCY LABELS

The proposed conventions are based primarily on a proposal by Arbisi-Kelm [3, 4] for transcribing aspects of American English speech (both typical and stuttered) [5], and was further informed by disfluency annotation systems for spontaneous speech in other languages [22, 28]. This proposal, as suggested in [12], specifies the annotation of individual disfluency-related acoustic cues; it provides separate markers for e.g. pause, prolongation, and cut-off words. Other disfluency phenomena, namely filled pauses, restarts or repetitions of words that are perceived as disfluent, are explicitly marked. The "prosodic phrase restart" (%r) label of existing ToBI conventions is maintained, and a new type of label for speech errors is included. These labels, shown in Table 1, can be marked either in isolation, or in combination with other cues. All cues are marked in a single tier, rather than being distributed across various tiers. In continuity with ToBI p-diacritic conventions, labels are used in the breaks tier. In contrast to previous conventions, however, the proposed conventions do not require the labeller to commit to a break index where cues to an intended well-formed target prosodic structure are absent.

Table 1: Proposed labels for disfluency phenomena

Label	Phenomenon & Description
c	cut: a partially-completed word
e	error: mispronunciation or wrong word
f	filled pause: filler words (e.g. *um, uh, mm*)
pr	prolongation: abnormal and/or incongruous lengthening of a segment within a word
ps **psw**	silent pause: perceived incongruous pause between (**ps**) or within (**psw**) words. (Note: **ps** label can be followed by **s** to mark the end of the silent interval.)
rs	restart word: repetition of a segment, word, or fragment (often after a word has been cut off)
%r	restart phrase: start of a new prosodic phrase after a previous phrase was not completed

3.1 Proposed conventions for use of labels

The following conventions for use of these labels in a time-aligned tier-based annotation system are proposed: 1) Labels are placed in the breaks tier. 2) Most labels (**c**, **pr**, **ps**, **psw**, **f**, **e**) are placed at the end (right edge) of the word interval, and where possible, with break indices. 3) Exceptions to these placement conventions are **rs** (restart word) and **%r** (restart phrase), which go at the left edge of the word interval, to indicate that what follows is part of a restart. The **s**, to indicate the end of a disfluent-sounding silence, likewise is placed at the word's left edge. 4) When more than one symbol is needed, use a period (.) as a delimiter (e.g. **1c.pr**, **1psw.ps**).

3.2 Comparing disfluency labelling schemes

This proposed disfluency labelling scheme makes two important distinctions not captured previously in ToBI. First, it separates prosodic disfluencies from morphosyntactic and segmental speech errors. A distinct label for an error, i.e. a wrong word or sound, is used independently of other disfluency labels which may or may not co-occur. Second, it separates various prosodic aspects of perceived disfluencies, giving more informative annotations, as characteristics co-occur variably. For example, we

may hear a disfluent pause after a word with or without prolongation; filled pauses with or without silent pauses; or a cutoff followed or not by a restart.

Figures 1 & 2 show examples that demonstrate some of the variability in the signal not captured using ToBI p-break annotation. Figure 1 shows a file labelled using existing conventions, including the p diacritic (tier 3), vs. using the proposed new cue-based annotations (tier 4). Here, tokens labelled 2p and 3p correspond to different disfluency cues and combinations: **3ps**, **1pr**, **3pr.ps**. Further, **pr** (prolongation) appears with a 4 break. While the use of the p-diacritic is expressly excluded from use with the 4 break, the **pr** label can be used with any break index, as disfluent-sounding prolongations can potentially occur in any prosodic position.

Figure 1: A labelled example with both p-breaks (tier 3) and new proposed labels (tier 4).

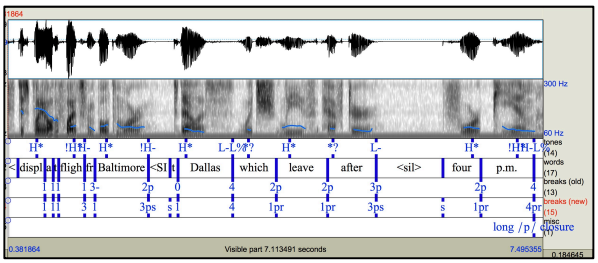
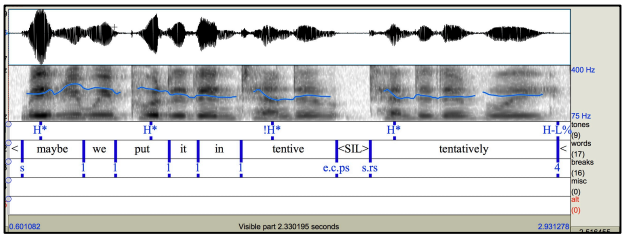


Figure 2 shows an example with a speech error. Existing conventions allow only the 1p label (for a cut-off word). The new labels reflect the multiple attributes of the token: a cut-off word (**c**), with pause (**ps**), that is also an error (**e**, the non-word "tentive"). The labels further indicate that the next word (*tentatively*) was a restart (**rs**).

Figure 2: A labelled example of a disfluency in which several cues and phenomena are captured.



4. SAMPLE DATA INVESTIGATIONS

To test the viability of this system, 2 expert ToBI labellers independently annotated disfluencies using the proposed labels of Table 1. Files labelled were 7 total files from 5 unique speakers from the American English Map Task (AEMT) database [2], collected in 1999 using the Map Task protocol [23]. Speakers were young adult female Boston area college students, from different US regions. (No further demographic data is available). Labelled files contained 7366 words and 1876 pauses.

4.1 Frequency of labels and label co-occurrence

We here present some brief observations about label frequency and co-occurrence in this data set. Labeller 1 tagged 1622 intervals with disfluency labels, and Labeller 2 1796 intervals, for a ratio of roughly 22 to 24 disfluencies per 100 words. Due to space constraints, data shown are from Labeller 1. (Future work will address inter-labeller differences; disagreement tended to relate to whether a cue, e.g. prolongation sounded disfluent.) Label distribution is summarized in Table 2. Individual word tokens could be labelled with no disfluency, or with one or more labels, e.g. just **pr**, just **ps** or **pr.ps** together.

Table 2: Frequency of label use by Labeller 1.

label	alone	in combi-nation	total	% of all disfluencies
c	85	147	232	14.3%
e	2	6	8	0.5%
f	61	127	188	11.6%
pr	453	287	740	45.6%
ps	329	324	653	40.3%
(s)*	453*	200	653*	n/a
psw	2	1	3	0.2%
rs	19	106	125	7.7%
%r	9	144	153	9.4%

* Occurrences of **s** (end of a silent interval labelled at start by **ps**), are not counted as separate disfluencies.

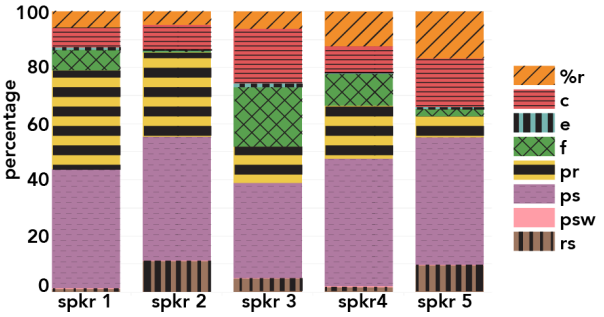
Of all disfluency-marked intervals, 960 were given a single disfluency label, and 662 some combination (18 with 3 disfluency labels, and the other 644 with 2 labels). Frequent 2-label combinations include **pr.ps** (prolongation and pause, 179 tokens), **c.ps** (cut-off and pause, 100) and **f.pr** (prolonged filled pause, 83). Other frequently co-occurring labels were **s.%r** (133 tokens) and **s.rs** (67 tokens), reflecting that prosodic phrase and word restarts occurred frequently after a disfluent pause.

4.2 Individual patterns of disfluencies

To illustrate the type of questions that this labelling system can shed light on, we examined inter-speaker differences in disfluency patterns [19]. Using the first file labelled from each of the 5 speakers, the frequency of each cue label was measured to create a disfluency profile. Disfluent pauses (**ps**) were the most frequent label used across these 5 files, while

other cues such as cut-offs (**c**) or filled pauses (**f**) varied more noticeably. Figure 3 illustrates the different disfluency patterns. This method also may shed light on intra-speaker differences and similarities across contexts. As we continue to label more files from each speaker, we may see that their patterns vary by context, time, or interlocutor.

Figure 3: Disfluency profiles of the 5 speakers.



5. GENERAL DISCUSSION

This prosodic cue-based labelling scheme reveals complexities in the realization of disfluencies in natural spontaneous speech, and has the potential to shed light on general speech planning processes, as well as on individual and group differences in disfluency production. To explore the potential advantages of this system for capturing variability, we plan to label more data from varied sources to investigate potential differences among groups, individuals and contexts. Our initial data set contains few speech errors, but other speakers or contexts may provide more errors. Our current data set is from a largely homogenous group of speakers; a more varied demographic group might allow investigation of the effects of aging, gender, dialectal differences, and mental state. (See [10] and [31] for discussions of factors affecting disfluency rates.)

In addition to speaker variation, we plan to further investigate cross-listener variability in perception. Might different individuals perceive some cues as disfluent, while others interpret them as well-formed timing variation? Initial examination of labeller differences suggests differing interpretations as to whether particular instantiations of cues, such as pause or prolongation, signalled a disfluency or a well-formed boundary. To better capture such variability in cue interpretation, and to reflect additional cue co-occurrence (such as filled with silent pauses), it may be useful to combine the labelling of disfluency cues with individual cues to other aspects of prosodic structure, including voice quality and f0 cues [14]. Brugos [11] suggests that f0 cues (e.g. whether there is reset, often seen with the %r label) may allow speakers to signal the intended interpretation of an utterance when temporal cues to prosodic structure are distorted by disfluency: "Speakers may use pitch to direct the listener about the proper interpretation of a disfluent utterance: continuous pitch across a disfluent pause or prolongation could be used to indicate continuation, where as a disruption of pitch continuity could therefore signal that the speaker has restarted." [11, p. 263] A cue-based scheme for labelling disfluencies as proposed here can also be integrated into other phonetic-cue-driven prosody labelling systems, such as RaP [16] or PoLaR [1].

As the system is further developed, refinements may prove useful, such as more fine-grained labels for prolongations [24]. It may likewise be fruitful to capture variation in filled pauses to reflect different segmental material [28, 24], which would be relevant to distinctions found in 'um' vs 'uh' in production and processing [15, 17]. Labels to better capture disfluent sequences and repair structure as in [35] or [30] will also be considered. Further, we hope to encourage dialog between investigators of linguistic prosody and investigators of disfluency phenomena from clinical perspectives. Labelling disfluencies together with aspects of prosodic structure may be beneficial to those working with spoken language in clinical contexts.

6. CONCLUSIONS

This proposal for labelling the individual correlates of speech disfluencies parallels developments in other speech-related domains, such as the labelling of individual acoustic cues to distinctive features [33], to prosodic boundaries [14] and prominences [1] and to individual kinematic characteristics of co-speech gestures [29]. Expanding the labelling system to a more fine-grained level in this way opens the door to discovering systematic relationships among cue patterns, meanings and linguistic context which are otherwise hard to discern. Labelling individual cues not only enables studies of how cues do and don't co-occur, it allows us to relate these patterns to the labels for well-formed prosodic boundaries that occur in the vicinity of a disfluency. Finally, we consider this method more user-friendly; relaxing the requirement to include intended constituent structure when cues to that structure are missing or distorted through disfluency. The labeller is relieved of the obligation to guess at speaker intent.

7. ACKNOWLEDGEMENTS

This research was in part supported by NSF grant #1451663.

8. REFERENCES

[1] Ahn, B., Veilleux, N, Shattuck-Hufnagel, S. 2019. Annotating Prosody with PoLaR: Conventions for a Decompositional Annotation System. *ICPhS 19,* Melbourne.

[2] American English Map Task (AEMT) database: https://dspace.mit.edu/handle/1721.1/32533

[3] Arbisi-Kelm, T. 2006. An intonational analysis of disfluency patterns in stuttering (Dissertation, UCLA).

[4] Arbisi-Kelm, T. 2010. Intonation structure and disfluency detection in stuttering, In C. Fougeron, et al (eds.) *LabPhon 10*. Berlin. 405-432.

[5] Arbisi-Kelm, T., Jun, S.-A. 2005. A comparison of disfluency patterns in normal and stuttered speech. Disfluency in Spontaneous Speech, Aix-en-Provence.

[6] Beckman, M., Hirschberg, J., Shattuck-Hufnagel, S. 2005. The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing*, OUP, Chapter 2, 9-54.

[7] Beckman, M., Ayers Elam, G., 1997. Guidelines for ToBI labeling, Version 3. Ms., Ohio State University.

[8] Beckman, M.E., Hirschberg, J. 1993/1997. The ToBI Annotation Conventions. Appendix A of the Guidelines for ToBI labelling.

[9] Beckman, M., Pierrehumbert, J. 1986. Intonational Structure in Japanese and English. *Phonology Yearbook 3*, 255–309.

[10] Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.F. and Brennan, S.E., 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2), 123-147.

[11] Brugos, A. 2015. *The interaction of pitch and timing in the perception of prosodic grouping*. Dissertation, Boston University.

[12] Brugos, A., Shattuck-Hufnagel, S. 2012. A proposal for labelling prosodic disfluencies in ToBI. Advancing Prosodic Transcription, Stuttgart**.**

[13] Brugos, A., Veilleux, N., Breen, M., Shattuck-Hufnagel, S. 2008. The Alternatives (Alt) tier for ToBI: advantages of capturing prosodic ambiguity. *Speech Prosody 2008*. Campinas. 273-276.

[14] Brugos, A., Breen, M., Veilleux, N., Barnes, J., Shattuck-Hufnagel, S. 2018. "Cue-based annotation and analysis of prosodic boundary events." *Speech Prosody 2018*, Poznan.

[15] Clark, H., Fox Tree, J. 2002. "Using uh and um in spontaneous speaking." *Cognition* 84, 73–111.

[16] Dilley, L., Breen, M. 2018. An enhanced autosegmental-metrical theory (AM+) facilitates phonetically transparent prosodic annotation. In *TAL2018,* 67-71.

[17] Fox Tree, J.E., 2001. Listeners' uses of um and uh in speech comprehension. *Memory and Cognition* 29, 320–326.

[18] Fraundorf, S., Watson, D. 2011. The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, 65(2), 161-175.

[19] Langston, A., Brugos, A., Shattuck-Hufnagel, S. 2018. Individual patterns of prosodic disfluency in Spontaneous American English Speech. ETAP4, Amherst, MA.

[20] MacGregor, L.J., Corley, M., Donaldson, D.I. 2009. Not all disfluencies are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language*, 111(1), 36-45.

[21] MacGregor, L., Corley, M., Donaldson, D. 2010. Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners. *Neuropsychologia*, *48*(14), 3982-3992.

[22] Maekawa, K., Kikuchi, H., Igarashi, Y., Venditti, J. 2002. X-JToBI: An extended J-ToBI for spontaneous speech. *ICSLP 2002*. Denver.

[23] McAllister, J., Sotillo, C., Bard E.G., Anderson, A.H. 1990. Using the map task to investigate variability in speech. Ms., Dept. Ling., Univ. of Edinburgh.

[24] McDougall, K., Duckworth, M. 2017. Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication* 95, 16-27.

[25] Nakatani, C., Shriberg, E. 1993. Proposal for labeling disfluencies in ToBI. Paper presented at the 3rd ToBI Labeling Workshop, OSU.

[26] Pierrehumbert, J., 1980. *The phonology and phonetics of English intonation.* Dissertation, MIT.

[27] Price, P., Ostendorf, M., Shattuck-Hufnagel, S., Fong, C. 1991. The use of prosody in syntactic disambiguation. *JASA*, *90*(6), 2956-2970.

[28] Rodríguez, L., Torres, I., Varona, A. 2001. Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish. DiSS'01, Edinburgh.

[29] Shattuck-Hufnagel, S., Ren, A. 2018. The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology*, 9.

[30] Shriberg, E., 1994. *Preliminaries to a theory of speech disfluencies.* Dissertation, UC Berkeley.

[31] Silber-Varod, V., Kreiner, H., Lovett, R., Levi-Belz, Y., Amir, N. 2016. Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals. *Speech Prosody 2016*, Boston.

[32] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. TOBI: a standard for labeling English prosody. *ICSLP-1992*, 867-870.

[33] Stevens, K. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *JASA* 111(4), 1872-1891.

[34] Veilleux, N., Brugos, A., Shattuck-Hufnagel, S., Patterson, A. 2007. "Distribution of Disfluencies and Errors in English Discourse," *ICPhS-07*, Saarbrucken. 1349-1352.

[35] Zayats, V., Ostendorf, M., Hajishirzi, H. 2014. Multi-domain disfluency and repair detection. *ICSA15.*

[1] The addition of 5th tier has also been proposed [13], the **alt** (for Alternatives) tier, to codify labeller uncertainty and/or ambiguity in the signal.