9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Speech, Prosody, and Machines: Nine Challenges for Prosody Research

Andrew Rosenberg

IBM Research
Yorktown Heights, NY, USA
amrosenb@us.ibm.com

Abstract

Speech technology is becoming commonplace. Traditional telephony based interactive voice systems have been joined by virtual assistants and navigation systems to create a broad ecosystem of voice enabled technologies. Prosody is an essential component to human communication, but machines still lag in their ability to understand information communicated prosodically and to produce human-like intonation.

This paper poses nine challenges designed to effectively and more thoroughly integrate prosody into current speech technologies. These include long-standing and contemporary concerns surrounding the availability and utility of data, gaps in linguistic theory and specific technological issues. Each of these challenges have received some attention, additional work is necessary to bring the role of prosody in speech technology closer to its role in human communication.

Index Terms: prosody, spoken conversation

1. Introduction

Speech technology is expanding into the public space at a remarkable pace. Prosody is a fundamental element of speech, and ought to be a fundamental element of speech technology.

Prosody research has made significant strides in explaining a wide range of prosodic phenomena both in terms of their acoustic realizations and their communicative functions. However, there are still significant gaps in our understanding of prosody, and techniques to leverage this information to speech technology. This paperposes nine challenges aimed to narrow these gaps. These challenges are motivated by the current state-of-the-art, though some are perennial and fundamental to prosody research more broadly, other are specific to current technological approaches and applications. While these three areas have significant overlap, the challenges are roughly organized around Data, Theory and Technology.

The Data: Regarding prosodic data, two limitations involve the amount of available data for investigation (Section 2.1) and in developing technology based using material from descriptive and domain-specific corpus studies (Section 2.2). There are a few approaches to generating more available prosodically labeled data. The first is to share more data. The second is to make annotation less resource intensive. The third is to algorithmically generate more data, data augmentation. There are some novel ways to address these traditional challenges. First, transfer learning, including "one-shot" learning, facilitates learning from one domain and leveraging these insights to another. This presents an opportunity to use small amounts of speech collected for a specific investigation in broader applications.

The Theory: The expanding range of speech applications is rapidly moving technology into scenarios where theory doesn't provide comprehensive guidance as to how information is communicated via prosody. There is a clear technological need to expand theories of prosody to describe the kind of speech that appears in dyadic conversations (Section 3.1). Second, prosody communicates multiple elements of information simultaneously. Most formal theory is under specified when it comes to describing how these aspects interact (Section 3.2). Finally, by increasing the users of speech technology, the number of non-native speakers necessarily grows. If a deployed system is overly dependent on native speech, its overall performance suffers. While there is good work on the prosody of emerging bilinguals, more is still needed to fully understand how these populations use prosody in communication (Section 3.3).

The Technology: Prosody research has had a significant impact in improving the naturalness of speech synthesis, and has found some successes improving information extraction from speech, speech assessment and extracting affect. However, speech technology does not broadly include prosodic information and there is still a gap between the information that humans and machines garner from prosody. Prosodic research can continue to make it easier to incorporate prosodic analysis into their systems. Tools need continued development to facilitate this. (Section 4.1). Speech synthesis poses two specific issues. First, the evaluation of prosodic assignment modules fails to account for the existence of multiple "correct" or equally natural productions of an utterance (Section 4.2). Second, end-to-end speech synthesis forgoes any explicit modeling of prosody. The challenges for prosody research here is to provide reliable prosodic control in an end-to-end framework, and to determine if this kind of framework can be used to promote understanding of the underlying prosodic phenomena (Section 4.3). Finally, a major challenge to prosody research is to demonstrate value to speech recognition (Section 4.4).

To summarize, the nine challenges are:

- #1 Make More Available Labeled Data
- #2 Bridge the Gap Between Experimental Data and Broadly Applicable Technology
- #3 Develop and Expand Pragmatic Theories of Prosody to Describe Conversational Phenomena
- #4 Develop and Expand Theories of Prosody to Describe Multiple Dimensions of Variation
- #5 Increase Understanding of the Prosody of Emerging Bilinguals
- #6 Lower the Barriers to Including Prosody in Applications
- #7 Improve the Objective Evaluation of Prosodic Assignment
- #8 Understand Prosody in End-to-End Speech Synthesis
- #9 Demonstrate the Value of Prosody to Speech Recognition

10.21437/SpeechProsody.2018-159

2. The Data

Data is the lifeblood of empirical research and modern technological development. High quality prosodic labeling is a resource intensive undertaking, therefore there is relatively little data available. Additionally, there is a tension between analysis of controlled speech and ecologically valid speech. Both pose difficulty for training broadly applicable technology.

2.1. Challenge #1: Make More Available Labeled Data

Prosody research is facilitated by labeled data. High quality systems for prosodic analysis or assignment require labeled data for training. Descriptive analyses and hypothesis testing require appropriate data, and frequently, some form of annotation. Some of this labeling takes the form of explicit and thorough prosodic transcription, like ToBI labeling [1]. Other labeling uses a reduced set of classes covering some subset of prosodic phenomena (like phrasing) or a more functionally inspired labeling, say, dialog act labeling, that is then used to investigate prosodic correlates. In most if not all cases, the cost of obtaining labeling data is quite high, limiting its availability. Regardless of what labeling is being used, and amount of material that is available, almost all researchers lament the amount of available labeled data. This shows up in statistical analyses where p-values "approach significance" and hopeful authors conjecture that this liminal result might in fact prove out if more examples where available. It also arises in modeling work where performance (in the main) improves as a function of the amount of training data.

Labeling data is expensive. Even skilled, careful labelers may disagree, resulting in requirements for quality control and redundancy. Training annotators is both costly and risky (due to attrition). This section offers four suggestions of what can be done to address this challenge.

Share more data There is not a lot of publicly available data that has been prosodically annotated. Notable exceptions include the Boston University Radio News Corpus [2], The Boston Directions Corpus [3], C-PROM (in French) [4], DIRNDL (in German) [5]. However, individual research labs, both in industry and academia, have annotated material under a constellation of annotation standards that are not shared broadly. There are a number of arguments offered for why this data isn't shared: *It is too valuable*. This may be valuable to the business, or valuable to the researchers if, say, the data has been collected and annotated, but there are still papers or dissertations yet to be completed based on it. *We're still collecting data, we're still cleaning the annotations*. We have a very limiting view that once a corpus is published that it is "final" and perfect. Although it has been used by many prosody researchers, the Boston University Radio News Corpus contains incomplete annotation, and a number of errors. As much as we hope they would be, annotations are almost never error free and data collections can always be expanded. *It is too small, who would be interested.* While possibly true, it is. *I don't want to open up the work to criticism.* This is troubling, but real. If an annotation is corrected or other errors found, a paper's findings can be undermined, and a research plan can be derailed. While we would like science to welcome criticism, there are real risks to the careers of researchers if their work is undervalued and dismissed. This is similar to the incentives and biases that can lead to statistically significant results being initially published, that subsequently fail to be reproduced [6].

I have previously advocated for a rethinking of how we treated corpora and their annotations [7]. By showing that some of the most famous corpora in Speech and NLP have errors, this work suggested that the community treat errors as the norm, and facilitate their correction. Version control as used in software engineering was proposed as a mechanism to do that. Reciprosody, developed by Reza, Rosenberg, Hirschberg, and Shattuck-Hufnagel, is a prototype web application that can provide an interface to an SVN backend to support this kind of support to share and maintain annotated corpora. For a number of reasons, this work has not yet resulted in a deployed and vibrant resource. However, the source code is available at `github.com/fahmidur/reciprosody`, and the needs for such a resource remain.

Reduce the Cost of Annotation It can take expert annotators up to an hour to annotate one minute of speech with full ToBI labels [8]. This does not include the cost of training an annotator, reconciliation of disagreements between annotators or labeler fatigue. To address this, Cole et alia developed Rapid Prosody Transcription (RPT) [9, 10, 11]. Under this approach, multiple listeners (between 10 and 20) assess the presence of prosodic phrase boundaries and prominences. Through a number of studies of this paradigm, RPT has been demonstrated that crowdsourcing is a viable platform to generate high quality annotations, at least of limited, and clearly describable phenomena. RPT's viability has been demonstrated on the most narrow the presence of prominence and phrasing. It remains to be shown that this kind of annotation can be extended to more nuanced distinctions, like contrastive stress or phrase ending types, but this should be taken as a viable proof of concept suggesting that high quality labeling might be available from non-expert participants. Marketplaces like AMT or Crowdflower might provide a sufficient pool of annotators assuming the annotation task is simple enough. While RPT is not likely to replace detailed, expert annotation, it can serve as a valuable approach to lower the cost of annotation, in situations where this degree of annotation is not necessary, or as a first pass annotation before more detailed annotation is complete.

Be inspired by Active Learning It's easier to collect data than it is to annotate it. It is common to believe that a corpus isn't useful until it is "complete". However, a significant amount of work is done on partially annotated corpora. This can be facilitated by active learning-like approaches to annotation. Active learning provides a framework to identify which data would be most valuable to annotate. Many active learning approaches involve training a classifier on a small amount of data, and using its predictions to generate a score describing how valuable a label for a particular instance would be [12]. This is typically characterized in terms of informativeness and representativeness of the unlabeled data. These qualities are related to the complementary characteristics of exploration and exploitation in reinforcement learning [13].

There is a good amount of work that discusses the merits of different strategies for active learning broadly. For prosody specifically, Fernandez and Ramabhadran showed that a variety of measures yield performance over a random baseline [14]. While there is nuance as to which measure is *best*, the differences are not all that large. This work has been reproduced in the prediction of phrase boundary in Chinese by Zhao and Ma [15].

When considering a new task, developing a full-blown active learning strategy may be out of scope. Incorporating machine learning into the annotation pipeline incurs a significant amount of overhead. Labs that perform data collection and annotation may not have appropriate expertise here. Even if it is available, is may be overkill for small collections. However, ap-

plying an active learning-*inspired* approach can be valuable for even moderately sized collections. There are approaches to assess the potential value of data that have lower resource requirements than classifier training. For example, in a spontaneous speech collection, it might be useful to start annotating speakers for whom there is the most data available. This can allow for a single speaker analysis to be performed prior to speaker independent studies. Alternately, it may be more valuable to annotate a small amount of data across many speakers. A quick analysis of data can omit utterances that are short, or contain no speech. Even simple approaches can speed up annotation efforts delivering more value quicker.

Algorithmically Create More Data Even with the approaches described above, research is limited by the amount of data available. Recently, Fernandez and Rosenberg investigated the use of label-preserving transformations for data augmentation [16]. This is a common activity in the vision community, where scaling, translation and rotation transformations are applied to images generating copies of the training data while preserving the label. A horse, is still a horse, even when the image is scaled, translated or rotated. For prosody, we need to identify transformations that maintain the prosodic content of speech while modifying other aspects. For this work, we looked to voice transformation to improve robustness to speaker differences. The GSVT voice transformation approach that we used maintained the duration of an utterance, but allowed for the transformation of pitch range, pitch scale, and voice quality. This algorithm can easily generate signals that are unintelligible. We identified seven configurations that sounded enough like human speech and maintained the prosody of the original speech. By increasing the training data by a factor of eight in this way, we showed that detection of prominence (pitch accent) and intonational phrase boundaries could be improved by a relative reduction of error of 4.75% and 8.74% in a speaker *independent* setting. In combination with out-of-domain speech from another corpus (data pooling), data augmentation reduced pitch accent detection errors by 8.15% and phrase detection errors by 6.89%. It remains to be shown that this data augmentation approach via prosody preserving transformations can be applied successfully to other prosodic analysis and synthesis tasks.

2.2. Challenge #2: Bridge the Gap Between Experimental Data and Broadly Applicable Technology

Prosody researchers need to make a decision between laboratory speech that may be prompted or elicited in some way or using found, naturally occurring, more ecologically valid [17], speech. Both have value in empirical studies, but it can be challenging to use the data collected in a lab to train models that are robust enough to naturally occurring speech. When using naturally occurring data, there can still be important differences between the source and content of the material in training a model to be used elsewhere.

Lab speech has its obvious merits for empirical studies, descriptive statistics, and hypothesis testing. However, controlling for confounds can create an unnatural setting for speech, leading to spoken material that is dissimilar from naturally occurring speech. Moreover, the speech collected in laboratory studies can be quite small, sometimes as few as tens of utterances. This can be sufficient to investigate a theoretical hypothesis. From a technological perspective, there is a significant challenge in incorporating these findings and this data into a trainable model fo some sort.

Typically machine learning algorithms require many examples for training. This can leave narrow lab-speech collections to be useful as examples more than training data. In this case, the best way to leverage findings from lab speech is as inspiration, guiding feature extraction and pointing to directions of meaningful variation that can be verified in a larger datasets that are closer to the material the application will be interacting with. However, in the last few years there has been an uptick in "one-shot learning" [18, 19, 20]. The goal of this research direction is to adapt a system trained on a large amount of data to model a new or related phenomenon with very little data. Much of this work is in the vision domain, but there are applications to voice conversion [21] and learning to synthesize novel words (segmental pronunciation) [22]. While there do not seem to be applications of this approach to prosody, this offers a possible route to leverage very few examples of, say, laboratory, speech into a broader application.

However, there are instances where elicited laboratory speech is more substantial. This is the case in a number of elicited speech tasks, including the AMI meeting corpus [23] and the Boston University Radio News Corpus [2], and the CSC Deceptive Speech Corpus (CSC) [24]. In these cases, the material is elicited in a way as to approximate how people would speak, but there are still some contrivances – artificial speaker roles in AMI, controlled incentives for deception in CSC, etc. In these cases, there is sufficient data to train a reliable model of prosodic phenomena. However, the domain difference between these corpora and naturally occurring speech remains an issue. When considering, for example, ToBI labeled speech, we can see clear corpus differences when training a speaker independent model to predict ToBI labels on one corpus and testing on another [25]. The two evaluated corpora are dialog speech from the Columbia Games Corpus [26] and monolog speech from the Boston Directions Corpus [3]. Average pitch accent detection performance shows a 5% relative increase in error due in cross corpus evaluation. Intonational phrase boundary detection is somewhat more robust, revealing a 2% degradation of f-measure in the cross-corpus condition. Rosenberg et al. [**?**] show similar results for phrase break assignment from text finding particular limitations in the assignment of phrase ending tones in a cross-corpus setting. The source of these errors could be differences between speaking styles, speaker identities, recording conditions, lexical interactions, or idiosyncrasies of the labelers who generated the gold-standard labeling.

Domain adaptation and transfer learning are machine learning approaches that seek to improve the performance of an existing model to a new domain or new task. Unsupervised adaptation requires some speech in the target domain, but no labels. Ananthakrishnan and Narayanan demonstrated the effectiveness of such an approach to improve pitch accent detection using either a language model or acoustic by 13.8% and 4.3%, respectively [27] Sanchez et al. showed that a similar approach can improve performance of emotion recognition, using lab speech to improve recognition of emotions displayed in 911 emergency calls with a 15% reduction of error [28]. Cuendet applied these techniques to sentence segmentation reducing the NIST error rate by 3.7% through including out-of-domain data [29]. These results suggest that while domain differences remain a concern, there are available approaches to limit the impact on overall performance.

Alternately a researcher can use totally found speech – material that is generated for some other purpose and analyzed for its prosodic content. Examples of this approach include analyses of TED Talks (e.g. [30]) and Supreme Court arguments

(e.g. [31]). Applying findings from this material may also suffer from domain mismatch, requiring transfer learning and other normalization similar to elicited lab speech. There are still further complications from using found speech. Naturally occurring speech is even more varied in terms of speakers, recording conditions, topics, speaking styles, etc. This requires additional levels of robustness to speaker and channel differences. In addition to the domain adaptation approaches described previously, there are some good approaches to promoting robustness to low-level feature extraction. The some approaches here include normalization of pitch and intensity contours, appropriate hyperparameter setting for acoustic analyses (e.g. [32]), and task driven feature modification like i-vectors, and its predecessor UBM-MAP and Joint Factor Analysis (e.g. [33]).

There is an additional issue in using found material. Researchers in prosody (and many other fields) are typically in edge cases and irregular phenomena rather than the realizations of common behaviors. This bias toward the uncommon is necessary to refine underspecified theories and promote a more thorough understanding. It's also a common trait of human cognition to overweight the importance (and the frequency) of uncommon events [34, 35]. This leads us place more value on understanding rare phenomena, i.e. edge cases, by overestimating their likelihood, and their importance. If a phenomenon occurs in 1% of utterances, it will require 100-times more found speech than lab speech to obtain material for investigation. This multiplicative factor increases again once quality control considerations come into effect. Is it necessary to analyze only a single speaker? Do utterances with disfluencies need to be omitted? How much control over phonetic context is necessary? It is not unreasonable to hypothesize that it might take orders of magnitude more data to match the conditions obtained in lab speech in a found context.

3. The Theory

This discussion of challenges in prosodic theory is driven by technological needs. Humans extract a wealth of information from the speech signal via prosody. Achievement of human-like speech understanding requires human-like assessment of prosody. Synthesis of human-like speech requires synthesis of human-like prosody.

Speech enabled technologies are becoming more and more commonplace. The broad application of conversational agents is driving technological needs in terms of prosodic analysis and appropriate synthesis. There are three aspects of prosodic theory that could provide valuable insight to applications. First, effective conversation involves a wealth of pragmatics and paralinguistics (Section 3.1). While the theory literature is fairly robust when it comes to the relationship between prosody and syntax and semantics, the relationship between prosody and pragmatics and paralinguistics has been less thoroughly explored. Second, formal prosodic theory largely eschews differing sources of prosodic variation in its description (Section 3.2). Since multiple orthogonal elements of information are communicated prosodically, it is challenging to disentangle their influences on prosodic contours and events. Third, conversational agents and other speech technology frequently engage with emerging bilingual speakers. These speakers demonstrate prosody that is influenced by their native language and the spoken language. This makes prosodic analysis of non-native speech uniquely challenging (Section 3.3).

3.1. Challenge #3: Develop and Expand Pragmatic Theories of Prosody to Describe Conversational Phenomena

There is a vast and varied array of information that is communicated prosodically. There are speaker identity qualities ranging from gender, accent/dialect/native language, age and other social signifiers. These are relatively stable over time and context. Speaker state characteristics are more transitory. These can include whether the speaker is congested, affect like anger, nervousness, and joy. Some personality traits (like charisma) are assessed in a context specific way to suggest that they are more like speaker-state than speaker-identity. At a still lower temporal resolution are utterance level effects covering the perlocutionary and illocutionary forces. These include information pertaining to dialog acts and discourse structuring and qualities like sarcasm and humor which have both pragmatic and paralinguistic qualities. Finally, we get to the most narrowly defined prosodic information: prominence, and phrasing. These impact syntax, and the semantic and pragmatic meaning of an utterance, by informing focus, information status, segmentation, syntactic attachment and scope. The ease in which lay people appreciate these qualities speaks to their ubiquity and necessity within human communication.

Pragmatic and paralingusitic information is essential in the kind of dyadic conversation that conversational agents engage in. Systems need to be able to assess a speaker's intention, what is being asked about, or for, whether a speaker has moved to a new topic, whether they are angry, content or confused. While there is evidence that this information is communicated, in part, via prosody, additional formal and functional theory is necessary to effectively describe how these qualities impact prosodic realization. A theory that covers conversational phenomena may be more complicated for three reasons: 1) the interaction with lexical content is difficult to disentangle. It is relatively easy to construct lexically ambiguous examples of broad and narrow focus. The relative importance of prosody and therefore its markedness may vary in lexically prescriptive (e.g. wh- questions) and ambiguous contexts (e.g. declarative questions) 2) conversational behavior spans utterances and interacts with turn-taking. Discourse structure frequently spans utterance boundaries. Much formal prosodic theory does not include a description of units larger than an intonational phrase. While analyses of information structure imply a discourse context that can be contributed to by an interlocutor, I am unaware of any formal theory that incorporates in its description dialog units that have been spoken by another speaker. However, this is the context in which a great deal of speech occurs (including speech with conversational agents.) 3) A functional connection to pragmatic meaning and paralinguistic content likely needs formal description at the contour level, even if component units are narrower. If a narrow description of prosodic events (as ToBI and IPO provide) is responsible for the formal link between intonational phonology and phonetics, analysis of contours will require a compositional description of the phonological units. This, of course, poses significant challenges to data collection by combinatorially increasing the number of examples that may be required for analysis.

Prieto [36] makes similar observations, writing "[t]he separation between the fields of intonational phonology and formal semantics/pragmatics has led to a lack of unified depiction of intonational meaning within the linguistic community.", while highlighting three persistent issues: 1) *compositionality* relating to the identification of primitives and their composition , 2) *duality of structure* teasing apart of linguistic and paralinguistic

meaning and phonological and phonetic qualities of intonation and 3) *context-dependency* how lexical and situational context impact intonational meaning.

While, there are still significant gaps in our understanding of how intonation impacts the pragmatic and semantic meaning in discourse and dialog, theory is moving in this direction. There are some formal theoretical frameworks regarding how prosodic aspects integrate with the meaning of an utterance ([37, 38]) and some more focused analyses under both Gricean [39, 40] and Post-Gricean [41] pragmatics. Moreover, Prieto highlights an array of research treating intonation as an "integral part of linguistic grammar" and typical findings that intonation is used to encode the *modal* aspects of propositions. These modal elements are essential to understanding conversation.

There is an opportunity for the technological community to support this as well. The data that is required to support this technology can be used in empirical and theoretical studies, even if it was originally collected for different purposes.

3.2. Challenge #4: Develop and Expand Theories of Prosody to Describe Multiple Dimensions of Variation

Prosody is simultaneously communicating information about speaker identity, speaker state, discourse effects and utterance meaning. These information streams are partly orthogonal. A speaker's identity may be reliably "normalized out" of an analysis of utterance meaning, via prominence and phrasing, say, to facilitate speaker-independent analyses. However, it's not clear how qualities like speaker state or discourse effects can (or should be) accounted for. For example, say, high arousal emotions are communicated via increased speaking rate and increased pitch range (e.g. [42]). But discourse segmentation is also indicated by pitch reset (resetting the pitch range to a wider setting) and increased speaking rate [43, 44]. This raises the question, when a speaker starts speaking faster, louder and at a higher pitch range, should we assume that they have changed emotional state, or introduced a new discourse segment? Moreover, can these various influences be described formally, or must we rely on a functional decomposition? This is related to Prieto's observations about compositionality and duality of structure [36].

This is a challenge for formal theories of prosody. Firstly, it blurs the formal/functional divide by desiring for a formal description of a functional decomposition. However, this is not without precedent. The Fujisaki superpositional model decomposes the influences of the phrase and accent commands [45]. A broader set of decompositions might make a formal analysis possible without prescribing a functional relationship or responsibility to each. However, does a formal description need to be superpositional or can a linear description accommodate the fact that prosody contains multiple informative signals?

There are two theoretical frameworks that are somewhat consistent with this concern. The Superposition of Functional Contours (SFC) model [38] takes an position that surface prosody is a realization of the superposition of contours each of which serve some communicative function. PENTA [46] is related to this, but rather than describing superpositional contours, this theory uses "prosodic encoding schemes" that are analogized to lexical morphemes via compositionality and allophonic variation. However, both of these are functional theories. One risk of relying on a functional theory of prosody comes from the mismatch between the functional classes and instantiations that are described by the theory and those that are required by the technology. It remains an open question as to whether or not a formal theory of prosody can address how multiple aspects of information are communicated prosodically.

The empirical work to develop and test these theories is resource intensive (cf. Section 2.1). Designing an experiment to investigate an interaction of two variables is multiplicatively more cumbersome than a single variable. To investigate a still broader intersection may require an exponential growth in conditions and data requirements. Logistically it would be helpful if such a theory were compatible with standards where there is already a wealth of available labeled data, like, ToBI. This would allow studies to leverage existing labeled data either for hypothesis testing or to serve as partial annotations.

3.3. Challenge #5: Increase Understanding of the Prosody of Emerging Bilinguals

Commercially deployed systems are all-but-guaranteed to interact with speech from emerging bilingual (or non-native) speakers. Worldwide, the majority of English speakers are not native speakers. The challenge here is understanding how the prosodic systems of a speaker's native language and spoken language interact. There is theoretical and empirical prosodic work on a wide range of languages. Even focusing only on ToBI, variants exist for dozens of languages [47, 48]. However, models are typically silent on how to describe speech produced by non-native speakers.

Some theory is critical here for a number of reasons. Emerging bilingual speech is not homogenous. The speaker's specific native language has an impact on their spoken production. Moreover, not all speakers with the same native language are equivalently fluent. The variation inherent in these two dependent variables here can quickly overwhelm even the most thoughtful empirical study.

There is a descriptive work that has been done to investigate the prosody in emerging bilinguals [49, 50, 51, 52, 53] and in bilingual code-switched speech [54]. A good deal of this work has a theoretical foundation [55], typically this is based around existing language acquisition theories to describe phenomena like language transfer [56]. However, formal theory would be bolstered by being able to describe the prosody of speakers who have learned a language later in life and may have limited proficiency in the use of the language's prosodic inventory. Technological applications that are sensitive to prosody will necessarily interact with emerging bilingual speech, and will need to be robust to its differences from native speech. Theory to promote understanding of what these similarities and differences may be would be especially helpful here.

4. The Technology

This section describes four specific technological challenges for prosody research: 1) Facilitating the incorporation and investigation of prosody into as many applications as possible (Section 4.1), 2) Evaluating prosodic assignment, where there are multiple "correct" answers (Section 4.2), 3) Remaining relevant to end-to-end models (Section 4.3), and 4) Demonstrating value to speech recognition (Section 4.4),.

4.1. Challenge #6: Lower the Barriers to Including Prosody in Applications

The inclusion of prosodic information into a system can incur costs to a research group of any size, whether a single grad student or a major technology company. Without certainty that proper assessment of prosody will improve performance, many

groups are (rationally) hesitant to make this investment. This barrier limits participation in prosody research by people in adjacent scientific and technical fields.

There is a lot of research demonstrating all of the information carried by prosodic signals and techniques for how to analyze and synthesize this. This makes adoption of prosodic analysis techniques attractive. The technological challenge for prosody research here is to lower the barrier for entry. There is a need for more and better tools that can analyze the prosodic content of speech. Two tools of note in this space are OpenSmile and AuToBI. Both of these are effective, but neither completely address this challenge.

OpenSmile is a toolkit that facilitates the extraction of acoustic features [57]. It has been extensively used in prosodic and paralinguistic analysis [58]. It operates through two levels of analysis, a set of frame based features termed "low-level descriptors" (LLD) and higher level "functionals" that are applied to these. The LLDs include features like pitch, noise to harmonic ratios, and MFCCs. The functionals include operations like mean, standard deviation, maximum, and quantile measures. Since all functionals can be applied to all LLDs, OpenSmile makes it very easy to extract thousands of acoustic features for exploration. This broad approach has merits, but it can be difficult for a less knowledgable user to know which of these measures will have value.

AuToBI is an toolkit that performs automated ToBI labeling [59]. It operates on an acoustic file, and optionally a description of word boundaries. If word boundaries are not available, AuToBI performs an acoustic-based pseudo-syllable segmentation and annotates these units. The feature extraction is performed internally, but the features used in classification can be exposed to the user. AuToBI treats ToBI labeling as six separate classification tasks: detection of pitch accent, intermediate phrase boundary and intonational phrase boundary, and classification of the tone markings associated with the three events. Classification is performed via the LIBLINEAR toolkit [60]. Source code is available at `https://github.com/AndrewRosenberg/AuToBI` including code for training new models. Pre-trained models are distributed at `http://eniac.cs.qc.cuny.edu/andrew/autobi/`.

While AuToBI and OpenSmile are useful tools for prosodic analysis and have both lowered the barriers of entry for an array of investigations still more work is necessary. For example, both tools would benefit interfaces to multiple programming languages. Additionally, AuToBI needs to incorporate recent RNN approaches to ToBI labeling [61]. Similar tools for other annotation standards would allow for broader hypothesis testing.

4.2. Challenge #7: Improve the Objective Evaluation of Prosodic Assignment

We understand that given a string of text there are multiple, equally correct ways to produce a sentence. This has implications for how we objectively evaluate performance of a prosodic assignment system. This argument holds whether the target is a categorical value (like a label of prominence or phrasing) or a continuous pitch target. The typical machine learning argument would be to keep a held out set of paired inputs and outputs $x_i, y_i \in X, Y$ that was not used during training. Evaluation of a hypothesized set of labels $\hat{y}_i = f(x_i)$ is compared to the ground truth y_i by some loss function $d(\hat{y}_i, y_i)$. There is a problem with this for prosodic assignment. The problem is that y_i does not represent every correct prosodic assignment given x_i, rather it is a single example drawn from a larger set Y_i representing every natural prosodic realization of x_i. There may be some unobserved $y_i' \in Y_i$ where $d(\hat{y}_i, y_i') < d(\hat{y}_i, y_i)$. This property means that the function that needs to be learned by some machine learning algorithm (or even hand crafted rules) isn't a function at all there is no $y = f(x)$ that can be learned. The target, y, is an element of a larger set.

This evaluation issue isn't unique to prosodic assignment. Similar relationships impact machine translation (there are multiple correct translations for any sentence) and document summarization (multiple summaries of a document are equally correct). For these two tasks, measures like BLEU [62] and ROUGE [63] have been developed to provide an objective evaluation measure that has a stronger correspondence with "quality" especially in the condition where there are multiple correct answers. Prosodic assignment (and speech synthesis more broadly) does not, to date, have a similar evaluation measure that is widely used.

One reason that this evaluation issue is more significant currently than in the past is the rise of neural networks that perform these tasks. Due, in part, to their nature as universal approximators, DNNs are prone to overfitting. A common training recipe involves monitoring the loss on a held-out, or development partition of the labeled data, and stopping training when the dev-set loss increases. However, when the element of the dev-set is a poor proxy for subjective quality, this can leave the model undertrained, resulting in less dynamic prosody. In some cases, we have observed that training a model beyond the minimal dev-set loss, a canonical case of "overfitting", results in more natural and expressive prosody. However, strictly minimizing training loss can result in actual overfitting, where the resulting quality is much worse. Without a reliable solution to this problem, we have not published a thorough description of these findings.

4.3. Challenge #8: Understand Prosody in End-to-End Speech Synthesis

End to end speech synthesis is accomplished by sequence to sequence modeling, embodied as an encoder-decoder RNN with attention [64, 65]. Such a model simultaneously learns an alignment from characters to audio and learns the mapping from representations learned from the character sequence to the specific acoustic representation. Typically these systems are trained on normalized text (with digits, acronyms and abbreviations expanded), but there is evidence that some text normalizations can be learned directly from data. The output features may be a spectrogram, or vocoder features. This kind of modeling obviates the need for most of the traditional synthesis frontend, eschewing explicit models for grapheme-to-phoneme conversion, linguistic feature extraction, and (critically to this discussion) prosody and duration models.

If prosody can be learned directly from (a sufficient amount of quality) text and paired (though not aligned) speech, is there still value of an explicit prosody model to speech synthesis? If the speech synthesis research community loses interest in the relationship between text and prosody empirical research activity in this space will be reduced. On the other hand, if such a model can advance the state of the art in prosody modeling, something can be learned through inspection of its operation. While neural models can be opaque, there is significant interest and activity in visualization and interpretability of their behaviors.

Currently, the baseline prosodic quality of end-to-end systems is quite natural, particularly on in-domain data. However,

there is not a clear mechanism for a user to control the prosody of a given utterance in an end-to-end synthesis framework. This poses a number of research questions regarding the relationship between input representations and resultant prosody. For example, do utterances need to be annotated with specific pitch and duration targets? Can abstract prosodic units (like ToBI labels, or emphasis) provide sufficient control? Can functional annotation, say, generated from a dialog manager, be used instead of formal annotation?

There has already been some work to improving the expressivity of end-to-end systems. Skerry-Ryan et al. [66] learn a fixed length embedding of prosodic content and find that they can transfer the prosody from one utterance to another by keeping this prosodic embedding constant. By conditioning on a prosodic embedding they find that a synthesized segment sounds significantly more like the reference utterance with target expressivity than an unconditioned baseline. However, they discovered that the prosodic transfer also impacted the perceived speaker identity. That is the resulting speech sounds more like the expressive style of the source expression, but it also sounds more like the source speaker. While understandable, this kind of interaction is not present in modular prosodic assignment modules. There is every reason to believe that this new approach to synthesis will continue to pose unforeseen challenges, particularly as it relates to prosody. Hopefully, in addition to improving synthesis quality, addressing these will expand our understanding of the relationship between text and prosody.

4.4. Challenge #9: Demonstrate the Value of Prosody to Speech Recognition

In speech synthesis, mapping from text to speech, prosody's value is well understood. Prosodic assignment, predicting prosody from text, is a major component of all synthesis systems and prosody is rightly considered an essential element of synthesis quality.

Speech recognition is the most visible and active applications of speech technology. When mapping from speech to text, the value of prosody has not been so clearly demonstrated. Most systems completely ignore prosody. However, most language acquisition research considers prosody to be an essential and primary component to acquiring speech recognition in humans [67, 68]. Empirical studies have shown clear connections between prosodic content and lexical content – suggesting that prosody is informative for recognition and disambiguation [69, 70]. There have even been a number of studies that have shown that including prosodic information in the acoustic model and language models improves performance [27, 71].

There are a few reasons that prosodic analysis has not become a fundamental component to speech recognition.

- *A historical explanation.* Speech recognition's early applications and successes come from recognizing words in isolation [72]. This is followed by recognition of a constrained vocabulary (e.g. [73]). until finally coming to Large Vocabulary Continuous Speech Recognition (LVCSR) [74]. Prosody will not be particularly informative to isolated word recognition and its value is still limited in the constrained vocabulary condition. Prosody can provide the most benefit in contexts where disambiguation via prominence, phrasing and even duration is beneficial. These include noisy conditions ad open-domain, conversational speech The disambiguating information communicated prosodically is more valuable to the pronunciation model and language model. However, information that could be used by these modules is generally discarded by the acoustic model. This decision simplifies processing by allowing the pronunciation and language models to operate on well understood categorical units, phones and words. This architecture makes it non-trivial to direct acoustic information to the areas of speech recognition that may benefit from it the most.
- *A domain explanation.* Prosodic signals, (i.e. intensity, pitch, duration) are, in general, more robust to noise and environmental effects. It is possible that more challenging acoustic conditions, where spectral information is less reliable, will show a greater benefit from prosodic signals. Recognition of multiple speakers in less controlled environments than telephony will also pose challenges to standard acoustic modeling that may be aided by the inclusion of prosodic signals.
- *A performance explanation.* While there are some studies that have shown that inclusion of prosody helps performance, these have been performed on relatively small datasets. Anecdotally, researchers have reported that these gains do not persist when more data is available. Most commercial systems leverage thousands of hours of transcribed data to train speech recognizers. Demonstrating performance gains on this scale is necessary.
- *A redundancy explanation.* Prosody might already be represented in some speech recognition systems. It is not uncommon for adjacent acoustic frames to be stacked before generating a posterior for a central frame. If this stacking is large enough, this may capture some suprasegmental context. As one example, the BUT Babel system stacked 21 10ms frames to as input [75]. A 210ms window of analysis certainly covers multiple phones and could capture some prosodic variation. Traditional acoustic models output a phone representation, but in GMM or Hybrid systems these categories are typically context-dependent (CD) phones whose identities are learned from data [76]. It is possible that these CD phones are capturing some correlation with the prosodic context of the phone in addition to the segmental context.
- *An evaluation explanation.* Speech recognition performance measured using Word Error Rate (WER). Critically, WER does not consider punctuation as a token to be evaluated. In addition to all of its other functions, a good deal of the information that prosody communicates is valuable for identifying punctuation in speech transcripts. However, ground truth transcripts of corpora like Switchboard [77] do not include punctuation. Therefore, systems aren't scored on their ability to correctly segment speech into sentences, or appropriately insert commas, quotation marks, or parentheses. I don't believe there to be a good justification for this decision. Both human and machine consumers of speech transcripts will benefit from punctuated text (e.g. [78]). It will, for example, make speech transcripts more like text material narrowing the difference between these genre for downstream SLP/NLP tasks. It will make transcripts more readable by humans. There is some evidence that this environment is changing. Google's speech to text API now includes punctuation for US English [79]. While there is no indication that prosody is being used to punctuate these transcripts, there is plenty of evidence to indicate that it would help [80, 81].

5. Conclusions

Speech technology has significant gaps in how prosodic information is analyzed, synthesized and understood. This paper

poses nine challenges to prosody research informed by the current state of the art in technology and theory to help close this gap. Some of the challenges, like Challenge #1: Make More Available Labeled Data, are long standing and have received quite a bit of research attention already, others are more recent, like Challenge #8: Understand Prosody in End-to-End Speech Synthesis. While there has already been important progress made on all nine, continued development in the coming years has the ability to make dramatic impact on speech technologies that are available to more and more people.

6. Acknowledgements

I am indebted to the scholarship of, and in more instances than I have deserved, conversation, guidance and mentorship from Julia Hirschberg, Mari Ostendorf, Bhuvana Ramabhadran, Raul Fernandez, Michael Picheny, Mark Hasegawa-Johnson, Elizabeth Shriberg, Mary Beckman, Kim Silverman, Janet Pierrehumbert, Hiroya Fujisaki, Stefanie Shattuck-Hufnagel, Sun-Ah Jun, and Jennifer Cole. A special thank you to Michael Picheny for comments on an early draft.

7. References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Proc. of the 1992 International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 12–16.

[2] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Boston University, Tech. Rep. ECS-95-001, March 1995.

[3] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: Studies on speech corpora," in *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[4] M. Avanzi, A. Simon, J.-P. Goldman, and A. Auchlin., "C-prom. an annotated corpus for french prominence studies," in *Proceedings of Prosodic Prominence: Perceptual and Automatic Identification, Proceedings of Speech Prosody 2010 Satellite Workshop*, 2010.

[5] K. Eckart, A. Riester, and K. Schweitzer, "A discourse information radio news database for linguistic analysis," in *Linked Data in Linguistics*, 2012.

[6] A. A. Aarts, J. E. Anderson, C. Anderson, P. Attridge, A. Attwood, J. Axt, M. Babel, v. Bahník, E. Baranski, M. Barnett-Cowan, E. Bartmess, J. Beer, R. Bell, H. Bentley, L. Beyan, G. Binion, D. Borsboom, A. Bosch, F. A. Bosco, and M. Penuliar, "Estimating the reproducibility of psychological science," *Science*, vol. 349, 08 2015.

[7] A. Rosenberg, "Rethinking the corpus: Moving towards dynamic linguistic resources," in *Interspeech*, 2012.

[8] A. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, "Automatic tobi prediction and alignment to speed manual labeling of prosody," *Speech Communication*, vol. 33, no. 1–2, pp. 135–151, January 2001.

[9] Y. Mo, J. Cole, and E. Lee, "Naïve listeners' prominence and boundary perception," in *Proceedings of the 4th International Conference on Speech Prosody, SP 2008*. International Speech Communications Association, 1 2008, pp. 735–738.

[10] J. Cole, T. Mahrt, and J. Hualde, "Listening for sound, listening for meaning: Task effects on prosodic transcription," *Proceedings of the International Conference on Speech Prosody*, pp. 859–863, 1 2014.

[11] J. Cole and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, no. 2, pp. 425–452, 11 2010.

[12] B. Settles, M. Craven, and L. Friedl, "Active learning with real annotation costs," in *In Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008, pp. 1–10.

[13] A. G. Barto and R. S. Sutton, *Reinforcement Learning*. Cambridge, MA: MIT Press, 1998.

[14] R. Fernandez and B. Ramabhadran, "Exploiting active-learning strategies for annotating prosodic events with limited labeled data," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic*, 2011, pp. 2208–2211. [Online]. Available: https://doi.org/10.1109/ICASSP.2011.5946919

[15] Z. Zhao and X. Ma, "Active learning for the prediction of prosodic phrase boundaries in chinese speech synthesis systems using conditional random fields," *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 1–5, 2015.

[16] R. Fernandez, A. Rosenberg, A. Sorin, B. Ramabhadran, and R. Hoory, "Voice-transformation-based data augmentation for prosodic classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 5530–5534. [Online]. Available: https://doi.org/10.1109/ICASSP.2017.7953214

[17] M. B. Brewer and W. D. Crano, "Research design and issues of validity," *Handbook of research methods in social and personality psychology*, pp. 3–16, 2000.

[18] J. Burgess, J. R. Lloyd, and Z. Ghahramani, "One-shot learning in discriminative neural networks," *arXiv preprint arXiv:1707.05562*, 2017.

[19] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "One-shot learning with memory-augmented neural networks," *arXiv preprint arXiv:1605.06065*, 2016.

[20] M. Scheutz, E. Krause, B. Oosterveld, T. Frasca, and R. Platt, "Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 1378–1386.

[21] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," *arXiv preprint arXiv:1802.06006*, 2018.

[22] B. Lake, C.-y. Lee, J. Glass, and J. Tenenbaum, "One-shot learning of generative speech concepts," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 36, no. 36, 2014.

[23] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[24] S. I. Columbia University and U. of Colorado Boulder, "Csc deceptive speech ldc2013s09," 2013.

[25] A. Rosenberg, "Modeling intensity contours and the interaction of pitch and intensity to improve automatic prosodic event detection and classification," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2012, pp. 376–381.

[26] A. Gravano, "Turn taking and affirmative cue words in task-oriented dialog," Ph.D. dissertation, Columbia University, 2009.

[27] S. Ananthakrishnan and S. S. Narayanan, "Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 138–149, jan 2009.

[28] M. Hewlett Sanchez, G. Tur, L. Ferrer, and D. Hakkani-Tur, "Domain adaptation and compensation for emotion detection," in *INTERSPEECH*, 01 2010, pp. 2874–2877.

[29] S. Cuendet, "Model adaptation for sentence unit segmentation from speech," 2006.

[30] M. Farrús, C. Lai, and J. D. Moore, "Paragraph-based prosodic cues for speech synthesis applications," in *Speech Prosody 2016*, 2016, pp. 1143–1147.

[31] Š. Beňuš, A. Gravano, R. Levitan, S. I. Levitan, L. Willson, and J. Hirschberg, "Entrainment, dominance and alliance in supreme court hearings," *Knowledge-Based Systems*, vol. 71, pp. 3–14, 2014.

[32] K. Evanini and C. Lai, "The importance of optimal parameter setting for pitch extraction," *The Journal of the Acoustical Society of America*, vol. 128, p. 2291, 10 2010.

[33] L. Ferrer, N. Scheffer, and E. Shriberg, "A comparison of approaches for modeling prosodic features in speaker recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4414–4417.

[34] R. Gonzalez and G. Wu, "On the shape of the probability weighting function," *Cognitive psychology*, vol. 38, no. 1, pp. 129–166, 1999.

[35] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," in *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013, pp. 99–127.

[36] P. Prieto, "Intonational meaning," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 6, no. 4, pp. 371–381, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1352

[37] H. Julia, *Pragmatics and Intonation*. Wiley-Blackwell, 2008, ch. 23, pp. 515–537. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470756959.ch23

[38] G. Bailly and B. Holm, "SFC: a trainable prosodic model," *Speech Communication*, vol. 46, no. 3-4, pp. 348–364, 2005. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00416724

[39] J. Hirschberg, "The pragmatics of intonational meaning," in *Speech Prosody 2002*, 2002.

[40] C. Gussenhoven, "Intonation and interpretation: phonetics and phonology," in *Speech Prosody 2002, International Conference*, 2002.

[41] D. Wilson and T. Wharton, "Relevance and prosody," *Journal of pragmatics*, vol. 38, no. 10, pp. 1559–1579, 2006.

[42] J. Tao and Y. Kang, "Features importance analysis for emotional speech classification," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 449–457.

[43] C. L. Smith, "Topic transitions and durational prosody in reading aloud: production and modeling," *Speech Communication*, vol. 42, no. 3-4, pp. 247–270, 2004.

[44] M. Swerts and R. Geluykens, "Prosody as a marker of information flow in spoken discourse," *Language and speech*, vol. 37, no. 1, pp. 21–43, 1994.

[45] H. Fujisaki and K. Hirose, "Modelling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation," in *Preprints of papers, Working group on intonation, 13th International Congress Linguists*, Tokyo, 1982, pp. 57–70.

[46] Y. Xu, "Speech melody as articulatorily implemented communicative functions." *Speech Communication*, vol. 46, no. 3-4, pp. 220–251, 2005. [Online]. Available: http://dblp.uni-trier.de/db/journals/speech/speech46.html#Xu05

[47] S.-A. Jun, *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press on Demand, 2006, vol. 1.

[48] ——, *Prosodic typology II: The phonology of intonation and phrasing*. Oxford University Press on Demand, 2014, vol. 2.

[49] N. G. Ward and P. Gallardo, "Non-native differences in prosodic-construction use," *Dialogue & Discourse*, vol. 8, no. 1, pp. 1–30, 2017.

[50] D. R. Verdugo, "The nature and patterning of native and non-native intonation in the expression of certainty and uncertainty: Pragmatic effects," *Journal of Pragmatics*, vol. 37, no. 12, pp. 2086–2115, 2005.

[51] D. M. Chun, *Discourse intonation in L2: From theory and research to practice*. John Benjamins Publishing, 2002, vol. 1.

[52] D. Ramírez Verdugo, "A study of intonation awareness and learning in non-native speakers of english," *Language Awareness*, vol. 15, no. 3, pp. 141–159, 2006.

[53] M. Cruz-Ferreira, "Perception and interpretation of non-native intonation patterns," in *Proceedings of the tenth International Congress of Phonetic Sciences*, 1984, pp. 565–569.

[54] D. Olson and M. Ortega-Llebaria, "The perceptual relevance of code switching and intonation in creating narrow focus," in *Selected proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology*, 2010, pp. 57–68.

[55] I. Mennen, "Beyond segments: Towards a l2 intonation learning theory," in *Prosody and language in contact*. Springer, 2015, pp. 171–188.

[56] J. Klassen, "Second language acquisition of focus prosody in english and spanish," Ph.D. dissertation, McGill University, 2015.

[57] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 2010.

[58] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *INTERSPEECH*, 2018.

[59] A. Rosenberg, "Autobi – a tool for automatic tobi annotation," in *Interspeech*, 2010.

[60] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1390681.1442794

[61] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling phrasing and prominence using deep recurrent learning," in *INTERSPEECH*, 2015.

[62] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[63] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL Workshop of Text Summarization*, 2004.

[64] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *International Conference on Learning Representations (Workshop Track)*, April 2017.

[65] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017.

[66] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *CoRR*, vol. abs/1803.09047, 2018. [Online]. Available: http://arxiv.org/abs/1803.09047

[67] S. R. Speer and K. Ito, "Prosody in first language acquisition - acquiring intonation as a tool to organize information in conversation," *Language and Linguistics Compass*, vol. 3, no. 1, pp. 90–110, 2009.

[68] L. Hahn, "Native speakers' reactions to non-native stress in english discourse," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1999.

[69] A. Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, D. Beaver, and D. Jurafsky, "To memorize or to predict: Prominence labeling in conversational speech," in *NAACL-HLT*, 2007.

[70] A. Rosenberg, "Using prominence and phrasing predictions to improve weighted dictionary pronunciation models," in *Interspeech*, 2012.

[71] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, "Prosody dependent speech recognition on radio news corpus of american english," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 232–245, 2006.

[72] K. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952.

[73] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 11, pp. 1870–1878, Nov 1990.

[74] G. Saon and J. T. Chien, "Large-vocabulary continuous speech recognition systems: A look at some recent advances," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, Nov 2012.

[75] M. Karafiát, F. Grézl, K. Veselý, M. Hannemann, I. Szöke, and J. Cernocký, "But 2014 babel system: analysis of adaptation in nn based systems," in *INTERSPEECH*, 2014.

[76] H. W. Hon and K. F. Lee, "On vocabulary-independent speech modeling," in *International Conference on Acoustics, Speech, and Signal Processing*, Apr 1990, pp. 725–728 vol.2.

[77] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, pp. 517–520 vol.1, Mar 1992.

[78] B. Favre, R. Grishman, D. Hillard, H. Ji, D. Hakkani-tür, and M. Ostendorf, "Punctuating speech for information extraction," in *In Proc. ICASSP*, 2008.

[79] "Google cloud speech-to-text updated w/ tailored video/phone models & auto punctuation," https://9to5google.com/2018/04/09/google-cloud-speech-to-text-updated-w-tailored-video-phone-models-auto-punctuation/.

[80] J. Kolár and L. Lamel, "Development and evaluation of automatic punctuation for french and english speech-to-text," in *INTERSPEECH*, 2012.

[81] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 5, pp. 1526–1540, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/TASL.2006.878255