Markers of confidence and correctness in spoken medical narratives

Kathryn Womack1, Cecilia Ovesdotter Alm2, Cara Calvelli3
Jeff B. Pelz4, Fengcheng Shi5, Anne Haake5

1 Dept. of ASL & Interpreting Edu., 2 Dept. of English, 3 College of Health Sciences & Tech.
4 Center for Imaging Science, 5 Computing & Information Sciences
Rochester Institute of Technology, Rochester, NY, USA
kaw8159@rit.edu, coagla@rit.edu, cfcsci@rit.edu
pelz@cis.rit.edu, spcast@rit.edu, arhics@rit.edu

Abstract
This study explores diagnostic correctness and physicians’ self-assessed feelings of confidence in spoken medical narratives. Dermatologists were shown images of dermatological cases and asked to narrate their diagnostic thought processes by providing a description of the case, a list of differential diagnoses, a final diagnosis, and, the percent confidence (from 0 to 100%) in their final diagnosis. We describe this novel corpus and present a case study from the dataset. We then report on predictive models for diagnostic accuracy and physician-reported confidence as a way of studying how these extralinguistic features affect the speech of the physicians, comparing use of narrative features, prosodic features, and disfluency features for classification.

Index Terms: spoken medical narratives, diagnostic correctness, diagnostic confidence, medical reasoning

1. Introduction
In the medical domain, the rate of diagnostic errors is estimated to be around 15% [1, 2]. Overconfidence is a significant cause of clinical misdiagnoses [1, 2, 3, 4, 5, 6]. Being able to recognize speech patterns of physicians who are confident versus unsure and correct versus incorrect in their final diagnosis would be useful, alongside eye-movement information [7], to better understand the cognitive reasoning processes of domain experts. This can serve as a first step toward mapping out where the diagnostic reasoning and decision-making processes break down when errors occur.

Specifically, there is a need for development of physician support systems and training systems which are user-centered and adaptive to the individual. We envision systems that can identify if users’ diagnostic reasoning strategies take a wrong turn, potentially leading toward an incorrect diagnosis. Tracking speech patterns that act as flags of decision-making problems or over- vs. underconfidence would be useful as one form of real-time behavioral sensors in such systems.

How uncertainty is affected by and plays a role in complex problem-solving is a prominent focus of current research. Highlighting the importance of accurately monitoring one’s self-perception of performance and actual performance, residents (experts in training) tend to show less correlation between self-reported confidence and diagnostic correctness [1, 8]. Physicisans who are more confident are also less likely to use decision support systems [9], which means that they could miss useful available information due to overconfidence. Systems that effectively contribute to training and accurate self-monitoring is one way to work toward reducing misdiagnoses.

We make several contributions in this work. First, we review the literature relating to confidence, both in and out of the clinical domain, and the development of medical expertise. Second, we present and analyze a novel dataset, comprising spoken narratives by dermatologists who were asked to narrate their medical reasoning process during medical image inspection tasks. Third, we report and compare predictive models that automatically infer diagnostic correctness or confidence using prosodic features, narrative features, and disfluency features. This allows us to better understand which features help predict confidence and diagnostic correctness respectively and better understand the confidence-correctness relationship.

2. Previous work
Characterizing how speech contributes to conveying confidence remains an unresolved topic, and describing the expression of confidence is a multi-faceted problem. Studies have looked at the listener’s perceived confidence as well as the perceptual rating by listeners of the speaker’s confidence. There is a lack of consensus on the relationship between actual versus perceived confidence and how they are expressed in speech.

Speech features play an important role in the way that listeners judge confidence. Intonation, the use of hedges or fillers, silent pause duration, and utterance duration have all been shown to inform automatic confidence classification or show statistical correlation [10, 11, 12, 13, 14]. Smith and Clark found in a question-answering study that respondents took longer and used more hedges when they were unsure of the correct answer [15]. This study considers medical decision making, which depends on expertise and reasoning as opposed to simple memory recall, but similar phenomena likely apply.

Importantly, judgments by listeners of speakers’ confidence also have effects on discourse and implications for learning and, accordingly, for designing learning systems. For example, Barr found in a teaching study that listeners learned more effectively from speakers who sounded confident (e.g. used few hesitations) and postulates that this is related to the listeners’ judgment of confidence in the speaker [16]. Moreover, Forbes-Riley and Litman found that a computer tutoring system that adjusted its response based on its perceived confidence of the human user resulted in more efficient learning by the student [17, 18], showing importance of both automatic confidence-detection and the importance of listeners’ perception of speakers’ confidence.

Although listener perception is important, many studies assume that listener perception accurately reflects the actual confidence level of the speaker, which is not necessarily the case.
Pon-Barry and Shieber report that, in their spoken question-answering study, self-reported confidence by participants was on average lower than the annotators’ rating of the speaker’s confidence [13]. Additionally, only 73% of their participants were labeled as “self-aware” ([13], p. 5), meaning that they were confident when correct and unsure when incorrect.

In the legal domain, facial recognition studies examining the accuracy and confidence of eyewitnesses have found weak or variable correlation between participants’ confidence and correctness [19, 20] and that the correlation can be changed by manipulating factors such as lighting and viewing duration [21, 22]. Similar findings hold true in the equally high-stakes domain of medical reasoning. Studies in both domains have shown stronger intra-subject correlation of accuracy and confidence and weak inter-subject correlation [8, 23], meaning that an individual might give consistent confidence ratings, but individual judgments of “high” and “low” confidence differ.

How decision makers’ confidence and diagnostic correctness interact in medical diagnostic decision making is also an ongoing subject of research. Many studies have found that there is weak or variable correlation between physician confidence and diagnostic accuracy (e.g. [1, 2, 24, 25]). Understanding how these two factors interact, which factors influence them, and how they are expressed in speech, would be useful to improve the human-centered design of physician training systems and, broadly, to further our understanding of how these extralinguistic characteristics are expressed.

The current study builds on previous findings by using a spoken corpus, in which we have both diagnostic correctness and self-reported diagnostic confidence ratings, in order to better understand how the two compare with respect to specific speech features. McCoy et al. postulated that indicators of speaker certainty help mirror the correctness of a medical expert speaker, with encouraging initial prediction results performing around or above tough baselines, despite a modest dataset size and notable class imbalance [26].

3. Data characterization and analysis

We use a novel corpus consisting of 8681 medical narratives, involving 29 dermatologists. 11 participants were practicing attending dermatologists and 18 were dermatology residents in training. Each participant was individually shown 30 images of dermatological conditions in a random order and asked to provide a description of the case; a list of differential diagnoses to consider; a final diagnosis; and a confidence score, a percentage from 0 to 100% of how confident they felt in their final diagnosis, with 100% being completely sure. The conditions included a variety of types of lesions and diseases that ranged from common to rare so as to include images in a range of difficulty.

All narratives are in English and were collected in an elicitation scenario in North America. Speech and eye-movements were simultaneously recorded; this study focuses on the spoken narratives, transcripts, and annotations. Each narrative was manually time-aligned by a single transcriber, with experience in this domain, using Praat [27]. All spoken tokens were transcribed, including silent pauses over 0.3 seconds and disfluencies (filled pauses, edits, etc.). A domain expert (co-author Dr. Calvelli) annotated narratives for diagnostic correctness.

The corpus contained over 92,000 tokens, with an average of about 100 tokens per narrative and approximately 1 minute duration per narrative. The average confidence score over all narratives was 63%. Boxplots comparing confidence scores and total narrative duration are in Figure 1. The mean confidence of participants on a specific image ranges from 44% to 98% (see images A and D in section 3.1); and physicians’ mean confidence scores over all images range from 25% to 84%. Table 1 shows the numbers of narratives with high vs. medium vs. low confidence, equally binned by score, and correct vs. incorrect diagnosis. More than two-thirds of correct narratives have a high confidence score. In addition, over three-quarters of incorrect narratives have a medium or high confidence score; estimating diagnostic accuracy appears in many cases to be problematic, with a tendency for overconfidence.

3.1. Case study

To provide an example of the narratives in this corpus and the effects of diagnostic correctness and participants’ confidence, we here examine four images, each having 29 narratives. Basic information about these four images is in Table 2 and images in Figure 2. Image A had the correct final diagnosis given by all participants, and the average confidence score was 98%. The pattern followed by most participants was to describe the lesion and immediately name the final diagnosis. Some participants listed one or two other possible diagnoses in the differential while others omitted the differential. This explains in part why the average duration of narratives for this image is shorter.

All participants gave the correct final diagnosis for image B, while none gave the correct final diagnosis for image C. Nonetheless, these two images have similar typical narrative structure patterns. For both, the average confidence score was high: 77%. Most participants in both images gave several pos-

---

narratives was 63%. Boxplots comparing confidence scores and total narrative duration are in Figure 1. The mean confidence of participants on a specific image ranges from 44% to 98% (see images A and D in section 3.1); and physicians’ mean confidence scores over all images range from 25% to 84%. Table 1 shows the numbers of narratives with high vs. medium vs. low confidence, equally binned by score, and correct vs. incorrect diagnosis. More than two-thirds of correct narratives have a high confidence score. In addition, over three-quarters of incorrect narratives have a medium or high confidence score; estimating diagnostic accuracy appears in many cases to be problematic, with a tendency for overconfidence.

3.1. Case study

To provide an example of the narratives in this corpus and the effects of diagnostic correctness and participants’ confidence, we here examine four images, each having 29 narratives. Basic information about these four images is in Table 2, with examples from spoken narratives in Table 3 and images in Figure 2. Image A had the correct final diagnosis given by all participants, and the average confidence score was 98%. The pattern followed by most participants was to describe the lesion and immediately name the final diagnosis. Some participants listed one or two other possible diagnoses in the differential while others omitted the differential. This explains in part why the average duration of narratives for this image is shorter.

All participants gave the correct final diagnosis for image B, while none gave the correct final diagnosis for image C. Nonetheless, these two images have similar typical narrative structure patterns. For both, the average confidence score was high: 77%. Most participants in both images gave several pos-

---

1Two narratives were lost in data collection due to technical issues.
There are two methods of clinical decision-making: the intuitive process that relies on exemplar-based pattern recognition, which is typically shorter; and the analytic process that steps through all available information, which is typically longer [6, 28]. The intuitive process is less reliable because it relies on heuristics and pattern-matching, and is therefore more subjective; whereas the analytic process takes longer but is more objective [28, 29, 30]. These case studies could show different patterns because of the distinctiveness of the diagnosis for image A, combined with the rarity of the diagnosis for image D.

Additionally, the diagnoses for images C and D rely in part on patient history and could be confirmed with clinical tests. The physicians are missing vital diagnostic information, and this explains the high rate of incorrect diagnoses, but this also provides us with valuable observations of how much information the physicians can (or cannot) glean from the image alone.

Many of the physicians seem to follow the intuitive process in image A, giving less evidence and a short (or omitted) differential. In contrast, for image D, many physicians seem to use the analytic process, discussing the case in depth and giving several options. Narratives in images B and C seem to be mixed, and it is not easy to tell exactly which method the physician might be using. These results provide additional evidence that physicians might resort to the more cautious analytic method when they are unsure or have insufficient experience to use the intuitive process, and that confidence and the dual reasoning methods are related. In addition, certain clues of spoken medical narratives’ structure (e.g. the narrative’s duration) could suggest which type of reasoning the physician is using.

4. Prediction models

The J48 decision tree in Weka\textsuperscript{3} was used for classification because of its ease of human readability and of evaluating individual features. Three groups of features, and their combinations, were evaluated at the narrative-level: prosody features, disfluency features, and narrative features, listed in Table 4.

For development, parameter tuning was performed with 60% of the corpus as a development training set and 20% as a development testing set. Following the tuning phase, given the small size of the corpus, we report on evaluation that used 15-fold cross-validation on the whole corpus. This partitioning ensured that training and testing sets differed compared to the development tuning phase and also allowed a more balanced understanding of the classification performance across test folds.

4.1. Predicting diagnostic correctness

Experiments\textsuperscript{4} for diagnostic correctness (i.e. correct or incorrect) were run for each of the feature group combinations, with and without the narratives’ confidence score as a feature. The majority-class baseline was 61%. Most of the experimental runs had above-baseline performance (see results in Table 5). Experiments that included confidence as a feature had an average of 5% higher accuracy than trials that did not include the confidence score. In fact, prediction using only confidence as a feature was as good as or better than other feature combinations (except, marginally, prosody with disfluency features and confidence), revealing confidence as a fair indicator of correctness. Decision tree inspection shows that very high confidence tends to be used to label correct diagnoses, but also results in substantial erroneous predictions, mirroring the complex link between confidence and correctness.

\textsuperscript{3}See http://www.cs.waikato.ac.nz/ml/weka/.
\textsuperscript{4}Parameters – confidence factor: 0.15, min instances to branch: 8.

### Table 2: For four images: the percent of participants who gave the correct diagnosis, mean confidence scores, mean narrative duration, mean number of tokens in a narrative, and mean number of diagnoses listed in the differentials.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100%</td>
<td>98%</td>
<td>42 s</td>
<td>75</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>100%</td>
<td>77%</td>
<td>58 s</td>
<td>101</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>0%</td>
<td>77%</td>
<td>62 s</td>
<td>103</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>0%</td>
<td>44%</td>
<td>92 s</td>
<td>140</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 2: Case study images. (Image C not shown for privacy.) Images used with permission from Logical Images, Inc.
Prosody features:
- Min. pitch (Hz)
- Min. intensity (dB)
- Time of min. pitch (s)
- Max. pitch (Hz)
- Max. intensity (dB)
- Time of max. pitch (s)
- Mean pitch (Hz)
- Mean intensity (dB)
- St. deviation of pitch
- St. deviation of intensity
- Pitch range (max - min)

Disfluency features:
- % Silent pauses (s)
- # Silent pauses / # tokens
- % Filled pauses (s)
- # Filled pauses / # tokens
- % False starts (s)
- # False starts / # tokens
- % Repetitions (s)
- # Repetitions / # tokens
- # Disfluent tokens / # non-disfluent tokens

Narrative features:
- # Tokens
- # Tokens per second
- Total duration (s)
- Total duration of speech (s)
- Type-token ratio (TTR)
- TTR (without disfluencies)
- Length of initial silent pause (s)

Table 4: List of features considered for classification. Features marked with * were useful for correctness prediction while features marked with + were useful for confidence prediction.

<table>
<thead>
<tr>
<th>P</th>
<th>N</th>
<th>D</th>
<th>P+N</th>
<th>P+D</th>
<th>N+D</th>
<th>All</th>
<th>ω</th>
</tr>
</thead>
<tbody>
<tr>
<td>+C</td>
<td>66</td>
<td>67</td>
<td>68</td>
<td>66</td>
<td>69</td>
<td>68</td>
<td>67</td>
</tr>
<tr>
<td>−C</td>
<td>60</td>
<td>63</td>
<td>61</td>
<td>62</td>
<td>60</td>
<td>64</td>
<td>64</td>
</tr>
</tbody>
</table>

| +Dx | 54 | 56 | 50 | 59 | 50 | 56 | 58 | 52 |
| −Dx | 52 | 57 | 52 | 55 | 51 | 56 | 56 | — |

Table 5: Top: Correctness prediction accuracy (%) with confidence as a feature (+ C) and removed (− C). Bottom: 3-scale confidence prediction accuracy (%) with correctness as a feature (+ Dx) and removed (− Dx). Results above baseline are in bold. P = prosody, N = narrative, D = disfluency features.

4.2. Predicting confidence

For confidence score prediction, we binned the confidence scores equally according to 2-scale, 3-scale, 4-scale, and 5-scale bins, to compare accuracy of each. Because the average confidence score was quite high, all trials had high majority-class baselines. All 2-scale prediction trials performed ±1% of baseline, likely because of class imbalance, as the majority-class baseline was 78%. Most trials for 3-scale and 4-scale predictions were above baseline, with 3-scale predictions (high, medium, and low; see Table 1) having the highest overall accuracy compared to its baseline, which was 52% (see Table 5).

Across the different scales for confidence, trials which included narrative features and correctness had the highest accuracy. This finding, that narrative features (such as duration) relate to confidence mirrors the discussion in section 3.1 that method of clinical decision-making (i.e. analytic vs. intuitive) and confidence also seem to be expressed via narrative features. The best performance in the 3-scale scenario was obtained with prosodic and narrative features, with diagnostic correctness (+7% above baseline). Including correctness as a feature improved most experiments, but not all, and correctness as the only feature did not score above baseline for the 3-scale case (in Table 5). This contrasts with the usefulness of confidence for predicting correctness, which could reflect self-awareness issues. The binning approach was ad-hoc and yielded unbalanced classes; alternative binning or sampling methods might capture more useful class boundaries.

4.3. Important features for classification

Weka’s CFS subset evaluator [31] was used after testing to see which features were valuable for classification. 15-fold cross-validation was used, providing the percentage of folds in which each feature was useful. A feature was considered important if it was useful in more than 50% of test folds (see Table 4).

For confidence predictions, the top features for each scale were narrative duration (total time the physician viewed the image) and duration of speech (from the time they started talking until they finished), which have shown differences in other analysis (see Figure 1, section 3.1). This explains in part why experiments involving narrative features out-performed other experiments. Correctness was useful for each confidence scale except 2-scale (but correctness as the only feature was uninformative, resulting only in majority class prediction, seen in Table 5).

Few features were useful for diagnostic correctness prediction; confidence score was the only feature that was useful in all of the testing folds. Useful features appear in each feature group, which could explain why there was low variation in accuracy between experiments.

Predicting diagnostic confidence and correctness, and comprehensively mapping out the confidence-correctness relationship, may benefit from more sophisticated algorithms and features, including semantic, clinical, or contextual features; as well as syntactic parsing or segmentation (e.g. [32]). Error analysis and other evaluation metrics, outside the scope of this work, may also yield new information.

5. Conclusion

We have described a novel dataset of medical narratives detailing dermatologists’ reasoning and decision-making processes while examining dermatological cases. Having physicians’ self-reported confidence in their final diagnosis provides a new facet for examining the relationship between confidence and correctness, and how they are encoded in spoken medical expert tasks.

Case studies of four images show a possible link with narrative-level features (e.g. narrative duration), diagnostic confidence, diagnostic correctness, and method of reasoning (i.e. analytic vs. intuitive). Automatic prediction classifiers for correctness and confidence show that confidence scores inform diagnostic correctness rather than the reverse, and that certain narrative features help predict confidence above baseline. The complex confidence-correctness relationship deserves more attention, e.g. how individuals’ self-awareness or decision-makers’ risk-taking profiles may play a role.

Results using shallow features are a good starting point, but highlight the need for deep-level meaningful features, including medical concepts [33]. This study has provided initial, tentative encouragement for predicting diagnostic correctness and confidence with some success, with potential for beneficial integration into multi-modal support systems that also consider other information and modalities, such as eye gaze.

6. Acknowledgements

This work was supported by NIH grant 1 R21 LM010039-01A1. We appreciate the contributions of the reviewers, participants, the transcriber, the research group; Dr. Lowell Goldsmith for helpful suggestions; and Logical Images, Inc. for images.
7. References


