PITCH-BASED EMPHASIS DETECTION FOR CHARACTERIZATION OF MEETING RECORDINGS

Lyndon S. Kennedy and Daniel P.W. Ellis

LabROSA, Dept. of Electrical Engineering, Columbia University, New York, NY 10027
lsk20@columbia.edu, dpwe@ee.columbia.edu

ABSTRACT

The automatic extraction of key utterances in spoken data has emerged as an interesting and difficult topic in automatic speech recognition. "Emphasis" or "excitement" may be a useful identifier for these utterances of interest. In this paper, we undertake the task of reliably and automatically identifying emphasized or excited utterances in natural speech in a meeting setting. We start by endeavoring to establish reliable ground truth emphasis labels by using several hand-labelers. The results show that human listeners can reliably identify emphasized utterances in meeting recordings. We then build an automatic emphasis detection system, which uses normalized pitch as its only acoustic predictor. The results show that this pitch-based emphasis detection scheme can distinguish between non-emphasized and emphasized utterances with an accuracy of 92% when ambiguous cases are excluded, a rate comparable to human interlabeler agreement.

1. INTRODUCTION

As we work towards making automatic speech recognition systems increasingly intelligent and natural, it becomes apparent that we not only want to be able to automatically understand which words the speakers have said, but also how they said them. Speech understanding systems based solely on the lexical content of speech data exclude the wealth of information conveyed by prosodic cues in natural speech. We instead wish to be able to extract the subtleties of the speaker's emotion as it is conveyed in the prosody of his or her speech.

There has been a significant amount of work in the area of extracting emotional and linguistic information from prosodic cues in speech signals [1, 2]. Much of this work, however, is based on extracting this information from recordings of prompted subjects acting out emotions, or from studies of humans interacting with automated systems. In this study, as in some other recent research [3], we work with natural human-to-human speech recordings and attempt to extract prosodic cues, namely speaker-normalized pitch, to indicate utterances with high emphasis or excitement.

In the next section we describe the meeting data we used for our experiments. Section 3 describes how we collected subjective ground-truth labels, then in sections 4 and 5 we describe how pitch data was extracted, and how that data was used for automatic emphasis labeling. Section 6 describes our evaluation of the automatic labeling, and section 7 presents the results. Discussion and conclusions are in section 8.

2. EXPERIMENTAL DATA

We developed and tested our system using the initial 22 minutes of a multi-channel recording of a meeting from a corpus prepared, transcribed, and labeled by ICSI [4]. The specific meeting was Bmr003, which is one of the 29 meetings consisting of different subsets of 8 participants who met regularly to discuss the Meeting Recorder project itself. (This excerpt had been used previously in evaluation of automatic speaker overlap detection [5].) Each of the participants is fitted with a high-quality, close-talking microphone. Additionally, there are 4 high-quality tabletop microphones and 2 lower-quality tabletop microphones. The meetings are hand-transcribed and include additional markings for microphone noise and human produced non-speech sounds (laughter, heavy breathing, etc.).

There are 6 speakers in the particular segment that we have used. 5 of them are equipped with head-mounted microphones and one is wearing a lavaliere microphone. We used the human-generated transcript to identify and segment the individual utterances from the speakers. The classifier relies entirely upon the close-microphone signals and ignores the tabletop recordings.

3. EMPHASIS LABELING

It has been shown that emotional perception is strongly influenced by context [6]. In some cases, context has been injected into the hand-labeling process by using labelers who personally knew the meeting participants, and thus were familiar with their speaking patterns [4]. In our labeling approach, however, we put the utterances into context for

the labelers by having the labelers mark ordered utterances from the same meeting. This allowed the labelers to hear each utterance in the context of the utterances that preceded it and in reference to the meeting as a whole.

To implement this labeling approach, we gave the labelers both a recording and a transcript of the meeting. The labelers listened to the meeting recording and followed along with the transcript, marking each utterance as "emphasized" or "neutral" as it was spoken. The labelers also had control over the playback of the meeting recording; they could stop and go back and review any portion of the meeting if they lost their place or needed to listen more closely to a particular section.

An interesting benefit of this method of generating hand labels was that it enabled the labelers to work almost in real time. Labelers reported that once a workflow had been established, they could label in sync with the flow of the meeting and rarely needed to stop or rewind. Labelers estimated that it took them each about 30 minutes to label the 22 minutes of speech data.

In the end, five undergraduate students at Columbia University labeled the 861 marked utterances in the 22-minute segment. We chose the labelers from a variety of disciplines and areas of interest (mostly not engineering or speech recognition) in order to better reflect the views of the average listener.

They were told to use their own best judgment when deciding which utterances were emphasized or neutral. They were not informed of the method that the automatic recognizer would be using when performing the same task, so they could not adjust their own criteria to comply with the automatic system.

All five labelers were in unanimous agreement on only 62% of the utterances. We took this data as permissible for the purposes of this study, however, for several reasons. Firstly, it does not appear that only one labeler is at fault for the lack of unanimity. When compared in pairs, each labeler agrees with each other labeler with comparable frequency (in the range of 75-85% of the time). Secondly, when we also count the cases where at least 4 out of the 5 labelers are in agreement, the percentage increases to about 84%. For these reasons, we can assume that the lack of uniformity in the labels is not due to incompetence on the part of the labelers, but rather simply due to the difficult nature of their task.

After we gathered the data from the five labelers, we used two methods to generate final, consensus labels for each utterance. In the first method, we used a simple majority decision. We chose the label chosen by 3 or more of the labelers as the correct label for the frame. In the second method, we required agreement between at least 4 labelers to confidently label an utterance as either emphasized or neutral. In the case where the vote was 3-2 (16% of the data), we labeled the utterance as "confusing." Confusing utterances were determined to be too difficult for humans to reliably classify, and thus the automated classifier was excused from having to analyze them: they were omitted from the error calculations.

Figure 1 summarizes the emphasis labels that were generated by each of the five labelers.

4. PITCH EXTRACTION

We used the Yin pitch estimator [7], which has recently been shown to be at the state-of-the-art [8], to perform the pitch extraction for the system.

We ran Yin over the six close-microphone signals from the meeting recording to extract pitch versus time for each of the speakers in the meeting. Yin returns two parameters of interest. Firstly, there is the actual pitch estimate, which is given as a deviation (in octaves) from A440 (440 Hz). One estimate is given for every segment of 32 samples. Secondly, there is the 'aperiodicity' which is a measure of just how aperiodic the signal is during a given sample. As a general rule, the more aperiodic the signal is, the less reliable the pitch estimate is. According to the Yin documentation, the pitch estimate is reliable when the square root of the aperiodicity is less than 0.3. This rule was used when deriving pitch estimates from Yin.

5. EMPHASIS DETECTION

The technique employed in this study builds upon the approach described by Arons [2]. Arons showed that emphasized segments in long single-person speeches could be extracted by locating segments of heightened pitch. He also showed that these emphasized segments tended to contain information that would be very helpful for generating summaries of the speeches.

In our system, Arons's approach is ported to the multi-speaker environment of a meeting, under the assumption that heightened pitch has a similar importance in acoustically characterizing speech data from meetings [3]. The speech signals of a meeting recording, however, differ considerably from the speech signals of a lecture recording, and a number of measures were taken to ensure that the pitch-based emphasis detection scheme transferred properly from the single-speaker mode to the multi-speaker mode.

Firstly, we noted that a baseline pitch must be determined for each of the speakers. Also, we must measure the pitch distribution for each speaker individually, since fluctuations of the same absolute size could mean very different things for different speakers. A speaker who uses a lot of pitch fluctuation in his or her regular speech would need to have an especially large increase in pitch in order to be considered as emphasizing his or her words. Likewise, a

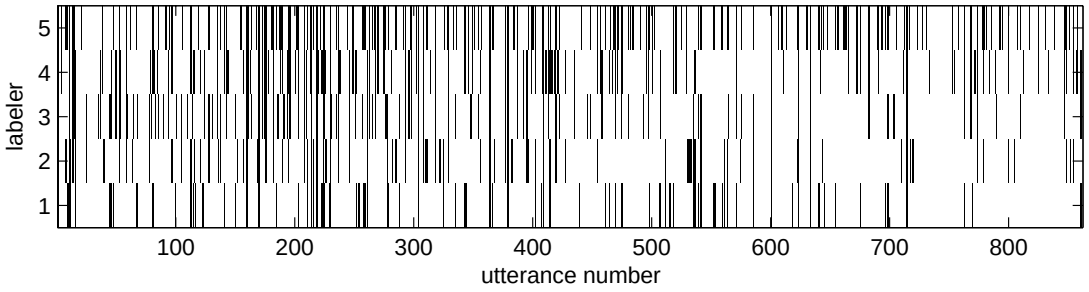

Fig. 1. Ground-truth of emphasized frames for each labeler

relatively monotonic speaker would not need to have such a large jump in pitch in order to be considered as emphasizing his or her words. Figure 2 shows the pitch distributions for each of the speakers in the meeting.

Secondly, a meeting setting raises the issue of segmenting speech not present in lectures. Lectures tend to be a single person speaking for many minutes at a time, while meetings are essentially a collection of short statements only seconds in length. Emphasized segments in lectures can be extracted by finding localized areas with high concentrations of emphasized frames, while emphasized segments in meeting recordings can be found by first segmenting the recording into utterances, which occur whenever a speaker becomes active, and then finding utterances that contain a greater-than-normal number of emphasized frames.

6. METHODOLOGY

We implemented the emphasis detection system by first extracting the pitch and aperiodicity for each frame on each of the close-microphone channels from the original recordings. We then calculated the mean and standard deviation for each speaker. To carry out this calculation, we applied two masks to the pitch values for each channel. The first mask cuts out any unwanted pitch estimates from the calculation by applying a threshold at square-root-of-aperiodicity values of 0.3 (as discussed in section 4). The second mask cuts out frames where the speaker is not actually speaking by choosing active frames from the transcript. After we applied the two masks, we were left with frames on each channel that contained the speaker appropriate to that channel uttering vowel sounds that had reliable pitch estimates.

We calculated the mean and standard deviation values for differently-sized training sets (5%,10%, 25%, and 50%), which were extracted chronologically from the available meeting data. The purpose for this is to explore the percentage of data necessary to reliably extract pitch pitch and emphasis distributions for different speakers. In situations in which multiple meetings need to be indexed on a rolling basis, it may be possible to determine a particular speaker's pitch

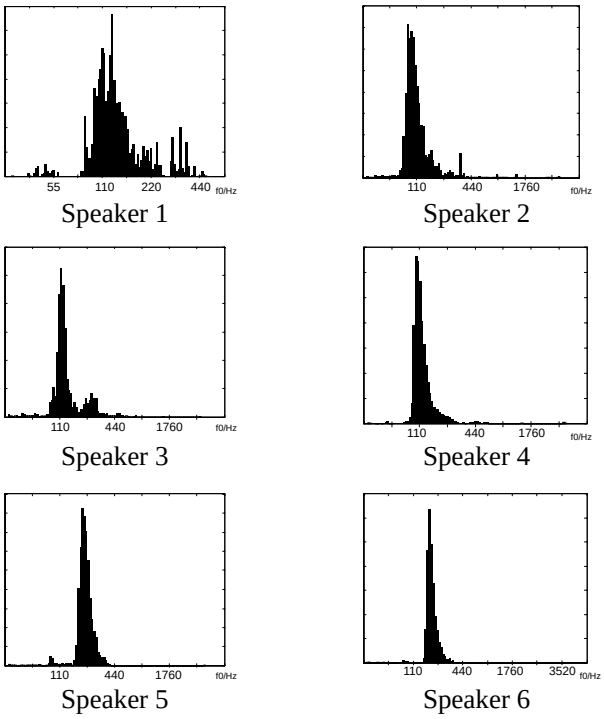

Fig. 2. Pitch distributions for each of the meeting participants.

distribution for only some of the meetings and then simply take that distribution to be constant and suitable for use in any future meetings.

Once we determined the pitch distributions for each speaker from the training data, we dynamically determined normalized thresholds to optimize the system's performance over the training data.

There were two thresholds being chosen: the first was to be applied to the normalized pitch deviation to detect heightened pitch on a frame level, and the second was to be applied to the percentage of frames with heightened pitch per utterance to determine utterance-level emphasis. We

Speaker	Mean Pitch	Pitch Std.	Emph %
1	111	0.48	0.08
2	104	0.35	0.16
3	122	0.44	0.21
4	119	0.37	0.13
5	205	0.42	0.04
6	181	0.29	0.05

Table 1. Mean pitch (in Hz), pitch standard deviation (in octaves), and emphasized frame percentages for each speaker.

Training Size	Precision	Recall	Accuracy
5%	8%	55%	72%
10%	17%	43%	85%
25%	16%	37%	86%
50%	19%	85%	90%
Random Guess	8%	13%	81%

Table 2. Summary of results including all utterances.

Training Size	Precision	Recall	Accuracy
5%	12%	43%	77%
10%	21%	55%	85%
25%	25%	33%	91%
50%	24%	73%	92%
Random Guess	7%	7%	89%

Table 3. Summary of results excluding "confusing" utterances.

	Detect Emph	Detect Not Emph
Emph	8	3
Not Emph	26	344

Table 4. Confusion matrix for the detection of the 11 emphasized segments and the 370 not-emphasized segments in the 50% training size case (excluding "confusing" utterances).

determined the thresholds by iterating through a range of values for each and comparing the results of the emphasis detection against the true hand-labels for the training set. We chose the pair of thresholds which maximized the harmonic mean (a standard measure of precision and recall) for the training set.

We then extracted the emphasized utterances from the testing set by applying the thresholds that we had determined with the training set. We first labeled all of the emphasized frames on each channel by finding frames in which the pitch deviates by more than the normalized threshold above the speaker's mean pitch. Secondly, we examined all of the utterances and labeled those that contain more than the normalized threshold above mean percentage of emphasized frames as emphasized utterances.

Table 1 gives each speaker's mean pitch (in Hz), pitch standard deviation (in octaves), and emphasized frame percentages.

Comparisons between the results of the scheme as described above and the hand-labeled emphasis information proved to be unsatisfactory. The classifier was giving an accuracy of about 53%. Qualitatively, the output was riddled with short utterances in which nothing significant (or emphasized) seemed to be said. What seemed to be happening was that there were a number of short utterances in the transcript in which the speaker on the channel only said something brief like, "yeah" or "um," but the actual signal on their microphone was dominated by whoever was actively speaking. So, there were situations where a male speaker with a low-pitched voice would mumble something under his breath while his female neighbor was speaking. The classifier ended up then detecting the high-pitched female voice and processing it as if the low-pitched male speaker was being very emphatic. Similarly, there were many utterances where the transcript contained a description of some microphone noise, but instead, a neighboring speaker dominated the actual signal on the channel.

Both of these problems were overcome by masking utterances of each type out of the signals used in the analysis. The first effect was eliminated by not looking for emphasis in any segment less than 1.2 seconds in duration. The second effect was eliminated by not using any frame in which the 'words' category of the transcript was contained entirely in braces (meaning that the utterance was only noise and contained no words from the speaker on that channel).

After these corrective masks were applied in the system, the results were much more satisfactory.

7. RESULTS

When a training size of 50% was used and the hand-labeled 'true' emphasized utterances were taken to be the utterances where a simple majority of the labelers agreed that the utterance was emphasized, the classifier performed with 19% precision and 85% recall for retrieving emphasized segments, and 90% accuracy for segment classification (which is strongly biased towards unemphasized). When the true emphasized utterances were taken to be only the utterances in which at least 4 labelers agreed that the utterance was emphasized and the utterances in which there was a split between the labelers were omitted from the classification as "confusing", the classifier performed with 24% precision, 73% recall, and 92% accuracy overall. The full results for all training sizes are summarized in Tables 2 and 3.

1.7 4: Oh there we go.
13.4 1: am up here *after a meeting,
348.3 3: Uh, for *other kinds of research,
369.9 3: I think I would *also very much like us to have a fair amount of *really random scattered meetings, of somebody coming down from campus, and -
469.4 4: Oh, oh, I'm not saying *accents.
473.0 4: No, it's more a matter of uh, *proficiency.
475.7 4: e- e- just simply *fluency.
481.6 4: undergraduates um in *computer science
488.5 3: Oh! You're not talking about foreign language at *all. You're just talking about -
659.8 4: Yes, that's fine.
758.8 1: Well, I know that space is really scarce on - at least in C_S.
781.8 1: Yeah, I think it would be interesting because then we could 'regularly get another meeting.
853.9 3: But on the other hand, it's not necessarily true that we need *all of the corpus to satisfy *all of it.
1103.9 1: The problem with engineers is "beep"
1120.0 5: I thought he meant, "Give them a music CD," like they g-
1127.8 5: you know, I personally would not want a C_D of my meeting, but

Fig. 3. The 17 utterances subjectively rated as most emphasized, presented as a kind of summary. First number is start time (in seconds), then speaker identity (1 to 5), then the hand-transcribed utterance.

25.2 1: been able to get that *error message in a point where I can sit down and find out where it's occurring in the *code.
331.0 5: constant or fairly similar, .. like a meeting about
440.7 4: O_K.
732.3 1: partner to do that .. we'd need to find someone on campus who was interested in this
787.0 1: type of meeting.
809.7 3: Yeah.
887.0 3: For th- for these issues of summarization, a lot of these higher level things you *don't really *need the distant microphone.
1050.9 3: Free lunch is good.
1062.9 3: {laugh}
1130.8 5: {inbreath}
1146.7 2: {laugh}
1164.6 1: I thought we could point that out.
1187.5 4: Well put, well put.
1238.7 1: still doing a bunch of archiving, I - I'm in the midst of doing
1262.9 1: the files,
1282.4 3: Is it?

Fig. 4. 17 utterances chosen at random, for comparison with the previous figure.

8. DISCUSSION AND CONCLUSIONS

As demonstrated by Tables 2 and 3, we must take care when considering the performance measures used to evaluate the system. The final measure of "Random Guess" listed in each table indicates the results that we see if we take the prior knowledge of how many true-emphasized utterances there are in each meeting and randomly select that number of utterances in the meeting as being emphasized. The percent accuracy of this method turns out to not be not significantly worse than the percent accuracy of our best-case detector. This is due to the presence of a large skew towards unemphasized utterances in the meeting data. Through the whole meeting, only about 15% of the frames are emphasized, so simply guessing that all frames are unemphasized will give the seemingly impressive result of 85% accuracy. Examination of the precision and recall measures, however, (also listed in Tables 2 and 3) reveals that the method presented is able to identify a much more significant percentage of the true emphasized utterances with significantly fewer false alarms than both random guessing or all-unemphasized guessing. So, indeed, there are significant benefits to using this analytical system over a scheme built on randomized or blanketed guesses.

Another unforeseen quality of the system we've presented is that the most emphasized utterances (as subjectively chosen by the hand-labelers) seem to give the qualitative appearance of a sort of summarization of the meeting's events. Figure 3 shows these most emphasized utterances. Someone who is familiar with the content of the meeting, perhaps one of its participants, should be able to recall many of the topics discussed in the meeting by looking at these utterances. The utterances that occur between 469.4 seconds and 488.5 seconds, for example, are from a rather involved discussion about what sort of speakers would be ideal candidates for participation as subjects in future meeting recordings. These utterances unambiguously signal that this discussion took place during this meeting. Figure 4, on the other hand, shows some utterances chosen at random from the entire meeting. It can easily be seen (even to someone who is familiar with the content of this meeting) that these utterances are significantly less indicative of the content of the meeting. The utterances which contain only noise descriptions or simple affirmative or negative responses are clearly not good indicators of meeting content.

Finally, this work also ties in closely with some recent work by Wrede [3] which attempts to determine whether or not "Hot Spots" can be identified on the utterance level by human labelers and presents a study indicating which features may be best for use in an automatic acoustic detector. ("Hot Spots" are identified as temporal locations in which multiple speakers are speaking with an increased emotional "involvement" in the conversation.) In our work, we have set out to determine whether or not "Emphasis" can be re-

liably identified on the utterance level by human labelers and we then hypothesize that it can be automatically extracted using speaker-normalized pitch as the only acoustic feature. Wrede finds that, of the many features that were examined, speaker-normalized pitch and energy are the best candidates for use as acoustic features in an acoustic detector of emotional involvement for the purposes of Hot Spot identification. In this paper, we have implemented a system using speaker-normalized pitch as the sole acoustic feature and seen that this detector can find emphasized utterances with accuracy and consistency on par with human labelers. From these results it seems likely that emotional "involvement" and "emphasis" are acoustically (and probably perceptually) very similar.

9. ACKNOWLEDGMENTS

This work was supported in part by the NSF under grant IIS-0212396 "Mapping Meetings."

10. REFERENCES

[1] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. ICSLP*, Denver CO, 2002.

[2] B. Arons, "Pitch-based emphasis detection for segmenting speech recordings," in *Proc. ICSLP*, Yokohama, 1994.

[3] B. Wrede and E. Shriberg, "Spotting 'hot spots' in meetings: Human judgements and prosodic cues," in *Proc. Eurospeech*, Geneva, 2003.

[4] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, , and A. Stolcke, "The meeting project at ICSI," in *Proc. HLT*, 2001, pp. 246–252.

[5] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *Proc. ASRU*, Italy, 2001.

[6] R. T. Cauldwell, "Where did the anger go? the role of context in interpreting emotion in speech," in *Proc. ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework*, Belfast, 2000, pp. 127–131.

[7] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, 2001.

[8] A. de Cheveigne and H. Kawahara, "Comparative evaluation of f0 estimation algorithms," in *Proc. Eurospeech*, Aalborg, 2001.