Addressee Detection for Dialog Systems Using Temporal and Spectral Dimensions of Speaking Style

Elizabeth Shriberg[1,2] *Andreas Stolcke*[1,2] *Suman Ravuri*[2]

[1]Microsoft Research, Mountain View, CA, U.S.A.
[2]International Computer Science Institute, Berkeley, CA, U.S.A.
{elshribe,anstolck}@microsoft.com, ravuri@icsi.berkeley.edu

Abstract

As dialog systems evolve to handle unconstrained input and for use in open environments, addressee detection (detecting speech to the system versus to other people) becomes an increasingly important challenge. We study a corpus in which speakers talk both to a system and to each other, and model two dimensions of speaking style that talkers modify when changing addressee: speech rhythm and vocal effort. For each dimension we design features that do not require speech recognition output, session normalization, speaker normalization, or dialog context. Detection experiments show that rhythm and effort features are complementary, outperform lexical models based on recognized words, and reduce error rates even if word recognition is error-free. Simulated online processing experiments show that all features need only the first couple seconds of speech. Finally, we find that temporal and spectral stylistic models can be trained on outside corpora, such as ATIS and ICSI meetings, with reasonable generalization to the target task, thus showing promise for domain-independent computer-versus-human addressee detectors.

Index Terms: addressee detection, dialog system, speaking style, prosody, language model, spectral tilt, vocal effort, online processing, out-of-domain data.

1. Introduction

As natural language understanding technology advances, speech input to dialog systems is allowed to be more free-form and conversational in nature. Speech interfaces are also increasingly used in open environments in which other people are present. Both factors conspire to make it necessary and challenging for systems to distinguish speech meant for the computer, from speech addressed to a human listener. We call this the addressee detection (AD) problem. It applies to both H-C dialog and to newer, H-H-C applications in which multiple people interact both with a system and with each other.

Past research on AD has focused on human-human (H-H) settings (such as meetings), sometimes with multimodal cues [1, 2]. Relatively little work has looked at the H-H-C scenario [3]. Early systems relied primarily on rejection of H-H utterances either because they could not be interpreted [4] or yielded low speech recognition confidence [5]. Some systems combine gaze with lexical and syntactic cues to detect H-H speech [6]. Others use relatively simple prosodic features based on pitch and energy in addition to those derived from automatic speech recognition (ASR) [7, 2].

In recent work on H-H-C data [8] we departed from past efforts and developed acoustic-prosodic features that do not require ASR, speech detection, or speaker and session-level normalization. *Independence from ASR* means that such features are more robust (since performance will not vary as a function of ASR errors) and can be deployed early in the processing pipeline. *Independence from speaker and session-level statistics* makes methods more robust to noisy environments and speaker variability, or movement of the speaker relative to the microphone. Assuming these constraints, we found that a signal-based measure of one dimension of speaking style, rhythmicity, is highly effective for the AD task. Follow-on work focusing on AD with word-based information [9] showed that lexical models can be trained on *out-of-domain* data, facilitating the development of systems in new domains.

This paper explores three new questions using a much larger data set than in [8]. First, *How can we capture a second dimension of speaking style in our data*—raised vocal effort—*using only the current utterance, without speaker or session information* (Section 2.3)? We explore voice quality features as well as local energy changes at voicing transitions, and investigate their effectiveness both alone and in combination with other features. Second, *How early do different acoustic-prosodic classifiers distinguish the addressee* (Section 3.2)? Here we analyze performance for *online* addressee detection for both acoustic-prosodic and lexical models in simulation experiments, to understand how they behave individually and in combination when given only the early portions of an utterance. Finally, *How well do acoustic-prosodic features generalize to completely different corpora that contain only one class (H-C or H-H)* (Section 3.3)? We train features on outside data selected for its addressee type, but otherwise substantially mismatched to our H-H-C data, and test how they generalize.

2. Method

2.1. Data

Data come from interactions between two acquaintances and a "Conversational Browser" (CB) dialog system using only spoken input. Subjects were brought into a room and seated about 5 feet away from a large TV screen and roughly 3 feet away from each other. They were told about the basic capabilities of the CB system and the domains it could handle, and were given a small set of short commands, such as to start a new interaction, pause, stop listening, or "wake up" the system. Other than that, subjects were told to use unrestricted natural language. The system detects starts and ends of utterances automatically. In this collection users did not use other modalities to indicate speech activity. More information about the dialog system itself and its spoken language understanding approach can be found in [10].

The resulting corpus comprises 6.3 hours of recordings over 38 sessions with 2 speakers each from a set of 36 unique speakers. Session durations ranged from 5 to 40 minutes, as determined by users. Speech was captured by a Kinect microphone

Table 1: Sizes, distribution, and examples of in-domain utterance types: H = human-directed, C = computer-directed, M = mixed.

	Train	*Test*
Utterances	2577	2889
Recognized words	7026	7874
H utterances	40.8%	31.0%
C-noncommand utterances	31.9%	32.8%
C-command utterances	24.5%	32.0%
M utterances	3.7%	4.2%

Type	*Example*
H	Do you want to watch a movie?
C-noncommand	How is the weather today?
C-command	Scroll down, Go back.
M	Show me sandwich shops. oh, are you vegan?

array; endpointing and recognition used an off-the-shelf recognizer. Although the full interaction was recorded, we used only the speech segments detected and recognized by the system; we simply call these "utterances."

A total of 6920 segments from 38 sessions were hand-transcribed at the word level, and annotated for addressee by an experienced experimenter who had run subjects in the data collection. The experimenter had access to the audio and video recordings and reported that annotation was generally straightforward. After eliminating utterances containing no intelligible foreground speech the dataset consisted of 5488 utterances, totaling 5.33 hours. Computer-addressed segments were also labeled by the annotator as either command or noncommand. Segments containing both human- and computer-addressed speech (in any sequence) were marked as mixed; since these were also processed by the system they were grouped with the computer-addressed class for detection purposes. The 38 available sessions varied greatly in length; the 22 shortest sessions were placed in the test set to maximize speaker and session variation. Table 1 gives the distribution of utterance types, and examples for each type.

For our experiments with out-of-domain training data in Section 3.3 we used a variety of corpora containing either human-human or human-computer speech. H-C corpora used included ATIS [11] and Communicator [12]; H-H corpora used were Fisher [13] and the ICSI Meeting corpus [14]. From each source, we sampled approximately 4 hours of speech data. For another contrast experiment, we used CB sessions collected with a single user, comprising about 2.8 hours of speech.

2.2. Lexical features

We used unigrams, bigrams, and trigrams of automatically recognized words ("**asr-ng**"), including start/end-of-utterance tags. The speech recognition system used had a word error rate of about 20%. For experiments to assess the best-case scenario for N-gram performance, we also extracted N-grams from human-produced reference transcripts ("**ref-ng**").

2.3. Acoustic-prosodic features

Energy contour DCT ("encon") features. To model rhythmicity we used a signal-based feature first proposed in [8], without any parameter tuning to the new corpus. The feature models the contour of 10-ms c0 and c1 output from an MFCC front end; each cepstral stream is mean-normalized over the utterance. A discrete cosine transform is then taken over a 200-ms sliding window with a 100-ms shift. Vector components comprise the first 5 and 2 bases from the DCT over each window of c0 and c1, respectively. Prior work modeled features with Gaussian mixtures (GMMs); here we also use a boosting model (see below).

Voice quality and spectral tilt ("tilt") features. H-C speech in our corpus tends to be produced with higher vocal effort than H-H speech. This likely reflects two factors: speech to a "less intelligent" (computer) listener, and speech to a distant microphone. As noted earlier, unlike previous work in AD, our challenge is to detect changes in vocal effort without using energy directly, since we have no way to normalize given our constraint of using only the current utterance. We thus looked at spectral-based measures [15, 16, 17, 18, 19, 20]. Features were extracted only for voiced regions. Voicing was determined using a binary decision for each 25-ms frame with 10-ms steps. A logistic regression classifier was trained with four features per frame: number of zero crossings, log energy, number of peaks in the autocorrelation of the window signal, and standard deviation of the inter-peak distance, using the Keele [21] and FDA [22] databases. The voicing threshold was always set to 0.5.

Under "tilt" features we include 5 measures. Three measures, H2-H1, F1-H1, F2-H1, use lower-order harmonics and formants. While [20] suggests higher-order formants are useful, they were not robustly identifiable in our far-field data. Similarly [16] and [23] suggest removing formant effects to correct for increased energy in harmonics; since we only extract lower harmonics, we did not pursue a correction. The last two measures are the spectral slope per frame (computed as the slope of a linear least squares fit to the log spectrum), and, following [19], the difference between the maximum of the log power spectrum and the maximum in the 2kHz-3kHz range.

Delta energy at voicing onsets/offsets ("devo"). We also explored whether raised vocal effort could be detected from the steepness of the energy slope at voicing onsets and offsets. Presently we use a crude estimation: the difference in log energy between frames centered around the onset/offset. We used collars of 1 through 5, 7 and 10 frames for both onsets and offsets, creating a 14-component feature vector for each utterance, modeled by boosting. Preliminary experiments showed that both onset and offset features are useful, so both types were retained.

Comparison features. Additional features were computed for comparison. With the exception of speaking rate measures, they are not expected to generalize well, but have been used in prior AD studies. Waveform length was included since a large percentage of H-C utterances are 1- or 2-word commands. Log energy was computed in voiced regions only; this was included to test how well raw energy performs compared to the tilt and devo measures. Voicing-related features include total voiced frames, the span of frames excluding initial and final pauses, the ratio of voiced to unvoiced frames, and statistics (mean, min, max, standard deviation) of durations of contiguously voiced regions. These were computed as a comparison to the energy contour feature, since they capture information about rate and rhythmicity. We also included a signal-based measure of speaking rate, "enrate" [24], as a comparison to the encon feature.

2.4. Classifiers and evaluation

We compute a log likelihood ratio of the two addressee classes from lexical N-grams by modeling each class with a **maximum entropy language model** (LM) [25]. Compared to earlier work on this corpus [8, 9], we found maxent estimation of N-gram probabilities to give consistent gains over traditional backoff language models. The detection score for an utterance w is computed as $\frac{1}{|w|} \log \frac{P(w|C)}{P(w|H)}$ where $|w|$ is the number of words in the test utterance, and $P(w|.)$ is computed by trigram LMs.

Table 2: System performance in %EER (g = GMM, b = boosting)

System	*EER*	*System*	*EER*
Individual features			
asr-ng	27.0	devo-b	26.2
ref-ng	10.4		
encon-g	16.8	tilt-g	33.8
encon-b	17.8	tilt-b	24.5
encon-(b+g)	15.4	tilt-(b+g)	21.7
System combinations			
encon-g + tilt-b			14.0
encon-g + devo-b			15.9
all-prosody = encon-(g+b) + tilt-(b+g) + devo-b			12.5
all-prosody + comparison features			12.6-12.8
all-prosody + asr-ng			10.9
all-prosody + ref-ng			7.5

The energy contour and tilt features employ **Gaussian mixture models** (GMM) to compute a log likelihood ratio. Training feature vectors for each class are pooled and a GMM with full covariances is trained. The score of a test utterance with feature vectors X then becomes $\frac{1}{|X|} \log \frac{p(X|C)}{p(X|H)}$ where $|X|$ is the number of vectors, and $p(X|.)$ is the aggregate GMM likelihood, assuming independence among the vectors.

Utterance-level features (both real-valued and binary) are modeled by the **adaptive boosting** algorithm [26] as implemented by [27]. Boosting induces a strong learner as a weighted combination of weak learners, each of which examines only a single feature of the input. The weighted combined classifier score then serves as a detection score in our experiments. For features that generate multiple vectors per utterance we first compute statistics (max, min, mean, standard deviation), both for the entire utterance, and over regions of contiguous feature vectors. Region-level statistics are then aggregated, yielding additional utterance level input features for boosting.

Linear logistic regression (LLR) was shown to be an effective combiner method for detection systems [28]. We use it to calibrate and combine one or more detection scores (obtained by any of the methods described above), yielding a single detection score per utterance.

Evaluation. To ensure calibrated comparisons, all systems, even those consisting of only a single model, were evaluated by performing LLR on their outputs. This was carried out by performing jack-knifing over all sessions in the test data, training the LLR parameters on all but one session, and cycling through all sessions. Scores are then pooled over the entire test set and evaluated using equal error rate (EER). The EER is obtained by choosing a decision threshold that equates false alarm and miss error probabilities. EER is thus independent of class priors. The behavior of AD systems is more fully characterized by a detection error tradeoff (DET) curve, showing how a moving decision threshold changes the two error type rates.

3. Results and Discussion

3.1. Acoustic-prosodic models

Table 2 shows EER performance of individual and combined systems, with the corresponding DET plot in Figure 1. The results for individual features (top part of Table 2) include combinations of GMM and boosting versions (b+g) for encon and tilt features (not included in the DET plot for space reasons).

Looking first at individual feature results, the noncheating word N-gram (ng-asr) performs worst overall. The cheating N-gram (ref-ng) is included for comparison purposes and gives best individual performance; this is largely because it gets 100%

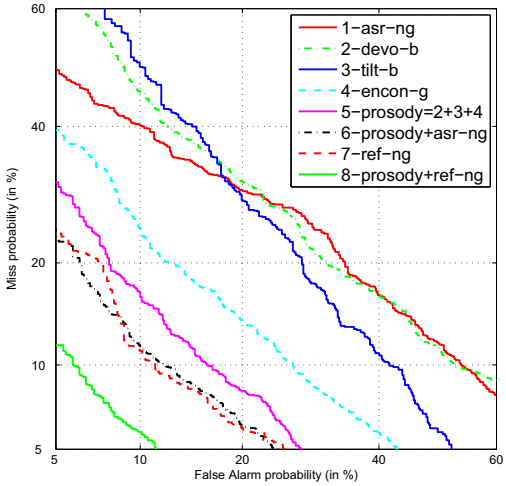

Figure 1: Detection error tradeoff

of commands (but not of longer queries) correct—a trivial task when commands are unique N-grams.

The best single prosodic system is the energy contour modeled by a GMM (encon-g). Interestingly the boosting version (encon-b) is close behind; this version uses only utterance-averaged statistics of the feature vectors. These two models using the same input features also combine well; we hypothesize that the boosting version is useful for shorter utterances while the GMM is more robust when enough window outputs are available. The encon features yield roughly half the error rate of the noncheating word N-gram.

We expected the vocal effort features to be less useful alone (because of the constraints on normalization and robustness of extraction issues discussed in Section 2.3), but to provide complementary information to the energy contour. This is indeed the case. As per Figure 1, tilt and devo features show different slopes; they combine well with each other even though both aim to capture raised voice. The first two results under "System combinations" also show that combining a boosting version of each of these features with encon gives gains over encon alone.

Voice quality features are best modeled via boosting (tilt-b). As we saw for the case of encon, there is some gain from combining GMM and boosting versions of the same features (tilt-b+g). Features measuring delta energy at voicing onsets and offsets (devo-b) perform only a little better than the asr-ng on their own. These features, however, are the only ones of many additional features we tried (including all of those listed as "comparison features") that improve performance once encon and tilt features are used. We combined the three prosody features (encon, tilt, and devo) into a single "all-prosody" system, which gives the low EER of 12.5%.

We asked whether the "comparison" features described in Section 2.3, modeled via boosting either all together or individually, could improve this result. These included utterance length, voicing pattern statistics, log energy in voiced regions, and the speaking rate measure "enrate". While each of these features performs well alone, none reduced the error rate of the 3-feature prosody system when added to or replacing one of the three features in the combination.

Adding the word N-gram to the 3-feature "all-prosody" system gives a further improvement. At an EER of 12.5% on its own, the prosody system is nearly as good as the cheating N-gram that has no errors on commands, as noted above. Interestingly, the prosody system provides a significant error reduction when added to the cheating N-gram; inspection shows that it re-

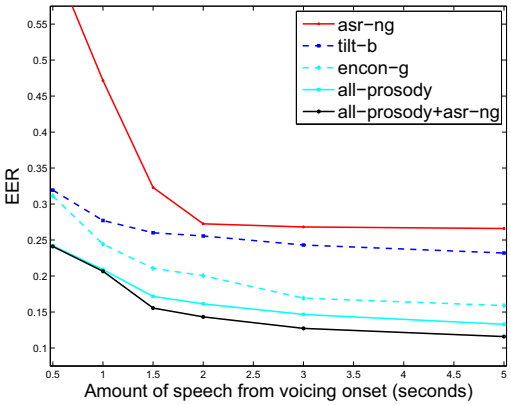

Figure 2: Simulated online AD performance

duces confusions between H-H speech and longer H-C queries.

3.2. Online processing

Figure 2 plots the performance of various detectors as a function of how much time from the beginning of an utterance is available to them. We thus simulate incremental online detection with increasing latency. Since many utterances have an initial pause, it makes sense to count time starting at the onset of speech. Since we did not want to rely on ASR, we chose the onset of the first voicing region as the origin of the time line. Separate analysis shows a good match between this estimated measure and the start of the first word (based on forced alignment). Note that encon features use only the available portion of the utterance for mean normalization.

All systems improve greatly over the first 2 seconds and then start to level off, in part because an increasing portion of all utterances are completed within the time window. The tilt system starts on par with the encon GMM; it is less good alone, but needs less initial data. The "all-prosody" system also includes tilt-g and encon-b systems (which are not shown individually). After 0.5 seconds, a combination of multiple prosodic systems improves over the best single system (encon-g) by a roughly constant amount at all time points. The same is true for the overall combination with asr-ng, except very early on in the utterance, where the lexical features provide little information.

3.3. Out-of-domain training

Finally, we are interested in whether acoustic-prosodic speaking style features generalize across different speech corpora. Since H-H-C data is currently scarce, we ask about generalization to corpora containing only one style (H-C) or (H-H). This makes the question interesting for both practical and theoretical reasons. We tested generalization using the best-performing of our stylistic models, the energy-contour GMM. Table 3 shows EER performance for a variety of training data sources described in Section 2.1. Systems were tested on the same in-domain data set as used earlier. "CBsingle" refers to training data from single-user sessions using the same system.

All out-of-domain sources give better than chance performance. ATIS and ICSI meetings were the best sources of H-C and H-H speech, respectively, out of our small sample of corpora. Jointly they give EERs that are only 31% relative higher than with in-domain training. Also, this combination of outside training sources beats any combination involving the single-user CB data, which is domain-matched to the H-C utterances.

Table 4 shows that the encon feature is the single best system trained on outside data—even better than the ASR-word-based model developed in [9] that is trained on all of CBsingle, Fisher, and ICSI meeting transcripts. As with in-domain data, we find that (a) tilt modeling, while less effective than encon, gives additional information, (b) combining multiple models based on the same features (GMM and boosting) helps in combination, and (c) prosodic and lexical models combine effectively, yielding an overall result that is only a few percentage points worse than the combined models trained on in-domain data. That tilt features provide any information is surprising, given that the ICSI meeting data is somewhat mismatched to our H-H data, and was collected in a large room and should therefore contain some raised vocal effort. Not shown in the table is that all "comparison features" performed poorly under mismatched training, lending support to our requirement that features not rely on information outside the current utterance.

Table 3: The encon-g performance for in- and out-of-domain training

H-C training	*H-H training*	*%EER*
CB (in-domain)	CB (in-domain)	17.3
CBsingle	CB	27.2
Communicator	CB	26.6
ATIS	CB	26.3
CB	Fisher	32.2
CB	ICSI	30.5
CBsingle	ICSI	27.8
ATIS	ICSI	22.8

Table 4: Various out-of-domain-trained systems and their combinations. All systems except asr-ng are trained on ATIS H-C and ICSI H-H data. The asr-ng system is trained on CBsingle H-C and Fisher plus ICSI H-H transcripts as in [9].

System	*%EER*	*System*	*%EER*
encon-g	22.8	tilt-b	40.6
encon-b	30.9	asr-ng	26.4
encon-g + tilt-b			20.3
encon-(g+b) + tilt-b			20.0
encon-(g+b) + tilt-b + asr-ng			15.9

4. Conclusions

We conclude that AD is enhanced by modeling both rhythmicity and vocal effort, using only signal-based features of the current utterance. These features outperform lexical addressee models and combine effectively with them; they also improve in combination with lexical models even if perfect word transcripts are available. Simulating online AD, we find that these features combine well with each other, and approach best performance given less than two seconds. Such early addressee detection could be used to reduce system latency and eliminating unnecessary processing. Finally, certain acoustic-prosodic features generalize far better than do word-based methods to out-of-domain data containing only H-C or H-H interaction—indicating consistency in speaking styles across domains and contexts. For practical purposes, results suggest that when limited H-H-C data is available for a particular application, one could use prosodic and lexical features to train effective AD classifiers entirely from out-of-domain, single-style data.

5. Acknowledgments

We thank Nelson Morgan for helpful suggestions on voicing transition features, and our colleagues M. Chinthakunta, A. Fidler, P. Greborio, D. Hakkani-Tür, L. Heck, G. Tur, P. Parthasarathy, and L. Stifelman for creating the infrastructure and data that enabled this research.

6. References

[1] R. op den Akker and D. Traum, "A comparison of addressee detection methods for multiparty conversations", *in Proceedings of Diaholmia*, pp. 99–106, 2009.

[2] N. Baba, H.-H. Huang, and Y. I. Nakano, "Addressee identification for human-human-agent multiparty conversations in different proxemics", *in Proceedings 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. ACM, Oct. 2012, Article no. 6.

[3] D. Bohus and E. Horvitz, "Multiparty turn taking in situated dialog: Study, lessons, and directions", *in Proceedings ACL SIGDIAL*, pp. 98–109, Portland, OR, June 2011.

[4] T. Paek, E. Horvitz, and E. Ringger, "Continuous listening for unconstrained spoken dialog", in B. Yuan, T. Huang, and X. Tang, editors, *Proc. ICSLP*, vol. 1, pp. 138–141, Beijing, Oct. 2000. China Military Friendship Publish.

[5] J. Dowding, R. Alena, W. J. Clancey, M. Sierhuis, and J. Graham, "Are you talking to me? dialogue systems supporting mixed teams of humans and robots", *in Proccedings AAAI Fall Symposium: Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems*, Washington, DC, Oct. 2006.

[6] M. Katzenmaier, R. Stiefelhagen, and T. Schultz, "Identifying the addressee in human-human-robot interactions based on head pose and speech", *in Proceedings of the 6th international conference on Multimodal interfaces*, pp. 144–151, State College, PA, USA, 2004. ACM.

[7] D. Reich, F. Putze, D. Heger, J. Ijsselmuiden, R. Stiefelhagen, and T. Schultz, "A real-time speech command detector for a smart control room", *in Proc. Interspeech*, pp. 2641–2644, Florence, Italy, Aug. 2011.

[8] E. Shriberg, A. Stolcke, D. Hakkani-Tr, and L. Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog", *in Proc. Interspeech*, pp. 334–337, Portland, Oregon, Sep. 2012.

[9] H. Lee, A. Stolcke, and E. Shriberg, "Using out-of-domain data for lexical addressee detection in human-human-computer dialog", *in Proceedings North American ACL/Human Language Technology Conference*, Atlanta, GA, June 2013.

[10] L. Heck, D. Hakkani-Tür, M. Chinthakunta, G. Tur, R. Iyer, P. Parthasarathy, L. Stifelman, A. Fidler, and E. Shriberg, "Multimodal conversational search and browse", *in Proceedings IEEE Workshop on Speech, Language and Audio in Multimedia*, Marseille, Aug. 2013.

[11] MADCOW, "Multi-site data collection for a spoken language corpus", *in Proc. DARPA SNP Workshop*, pp. 7–14, Harriman, NY, Feb. 1992. Defense Advanced Research Projects Agency, Information Science and Technology Office.

[12] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker, "DARPA Communicator dialog travel planning systems: The June 2000 data collection", in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001.

[13] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: a resource for the next generations of speech-to-text", *in Proceedings 4th International Conference on Language Resources and Evaluation*, pp. 69–71, Lisbon, May 2004.

[14] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus", *in Proc. ICASSP*, vol. 1, pp. 364–367, Hong Kong, Apr. 2003.

[15] J.-S. Lienard and M.-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels", *Journal of the Acoustical Society of America*, vol. 106, pp. 411–422, 1999.

[16] H. Hanson, "Glottal characteristics of female speakers: Acoustic correlates", *Journal of the Acoustical Society of America*, vol. 101, pp. 466–481, 1997.

[17] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children", *Journal of the Acoustical Society of America*, vol. 107, pp. 3438–3451, 2000.

[18] N. Obin, "Cries and whispers – classification of vocal effort in expressive speech", *in Proc. Interspeech*, Portland, Oregon, Sep. 2012.

[19] S. Takeda, Y. Ueno, N. Nakasako, H. N. M. Tsuru, R. Isobe, and S. Kiryu, "Spectral-tilt features of emotional speech", *in International Conference on Kansei Engineering and Emotion Research*, Paris, Mar. 2010.

[20] Y.-L. Shue, *The Voice Source in Speech Production: Data, Analysis and Models*, PhD thesis, University of California Los Angeles, 2010.

[21] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database", in J. M. Pardo, E. Enríquez, J. Ortega, J. Ferreiros, J. Macías, and F. J. Valverde, editors, *Proc. EUROSPEECH*, pp. 837–840, Madrid, Sep. 1995.

[22] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching", *in Proc. EUROSPEECH*, pp. 1003–1006, Berlin, Sep. 1993.

[23] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source", *Journal of the Acoustical Society of America*, vol. 121, pp. 2283–2295, 2007.

[24] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate", in G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proc. EUROSPEECH*, vol. 4, pp. 2079–2082, Rhodes, Greece, Sep. 1997.

[25] T. Alumäe and M. Kurimo, "Efficient estimation of maximum entropy language models with N-gram features: An SRILM extension", *in Proc. Interspeech*, pp. 1820–1823, Portland, Oregon, Sep. 2012.

[26] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization", *Machine Learning*, vol. 39, pp. 135–168, 2000.

[27] B. Favre, D. Hakkani-Tür, and S. Cuendet, "icsiboost. open-source implementation of Boostexter", http://code.google.come/p/icsiboost, 2007.

[28] S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the NIST'99 1-speaker submissions", *Digital Signal Processing*, vol. 10, pp. 237–248, Jan. 2000.