Prosodic Scoring of Recognition Outputs in the JUPITER Domain

Chao Wang and Stephanie Seneff

Spoken Language Systems Group, MIT Laboratory for Computer Science
200 Technology Square, Room 639, Cambridge, MA 02139, USA
{wangc,seneff}@sls.lcs.mit.edu

Abstract

JUPITER is a conversational system that allows users to access weather information over the telephone using natural speech [1]. This work examines the use of prosodic information to predict speech recognition errors more accurately for improved system robustness. Two approaches were explored here. The first approach is based on a probabilistic confidence scoring framework, which uses prosodic cues as additional features to improve both utterance-level and word-level confidence scoring. The second approach aims at scoring part of the prosodic space, focusing on phrases that bear important communicative functions. We explored the feasibility of characterizing directly the F_0 contours of some carefully selected English phrase patterns. We envision that these models can be applied to resort recognizer N-best outputs or to support rejection.

1. Introduction

This paper explores the use of prosodic aspects of speech to aid in the recognition and understanding of spontaneously spoken utterances. The research is conducted within the framework of the JUPITER weather-information domain, mainly because we have available a large corpus of utterances collected from natural telephone dialogues with a conversational system. The first half of the paper concerns the use of typical prosodic features, mainly fundamental frequency of voicing (F_0), energy, and duration, to improve confidence scoring, both at the word and at the utterance levels. The second half explores the much less well-defined area of higher level prosodic contours, and attempts to formulate concrete methods to utilize them in assessing the plausibility of a hypothesized utterance. We formulate a methodology, and attempt to discover any reliable patterns in the F_0 contour.

2. Confidence Scoring

Prosodic information can potentially assist in confidence scoring for several reasons. Hyperarticulated speech [2], which is associated with a slower speaking rate and increased F_0 and loudness, is likely to lead to degradation in speech recognition performance [3]. Furthermore, there are several prosodic cues to speech artifacts. For example, background speech from extraneous talkers is likely to be much weaker than the conversant's speech. Finally, incorrect hypotheses can exhibit anomalous prosodic aspects due to such obvious errors as an inappropriate stress pattern. We anticipate that "unusual" prosodic measurements will thus be indicative of speech recognition errors.

We have observed that utterances with a high percentage of internal silence are more likely to be incorrectly recognized. The internal pauses are usually associated with hesitation, emphasis, or hyperarticulation. Utterances with high mean F_0 are also more likely to be incorrectly recognized. This is consistent with the recognition results that female and child speech have considerably higher error rates.

We will first introduce previous work done by Hirschberg and colleagues on using prosodic cues in utterance-level confidence scoring. We then describe the confidence scoring framework used in our experiments. Finally, we report the utterance-level and word-level confidence scoring experiments in detail.

2.1. Related Research

The idea of predicting speech recognition performance from prosodic cues has been explored by Hirschberg *et al.* [4, 5] on a couple of recognition systems and application domains. Eight prosodic features were examined as potential cues to predict system errors in recognizing or understanding each user utterance. These features include maximum and mean F_0 values, maximum and mean energy values, total duration, length of the pause preceding the turn, number of syllables per second in the turn (tempo), and percentage of silence within the turn. Statistically significant differences were found in the *mean* values of a subset of these prosodic features between correctly recognized and misrecognized user turns. A *rule-based* classifier performed accept/reject decisions on recognition outputs, in conjunction with other information such as acoustic confidence score, language model, recognized string, likelihood score, and system prompt.

The results suggest that the efficacy of prosodic features depends highly on the quality of the recognition system. In the system which used "older" recognition technology and "poorer performing" acoustic and language models, the prosodic features achieved a large improvement over using acoustic confidence alone (over 50% reduction in classification errors), and the best-performing rule set included prosodic features. However, in the system which was better trained for the recognition task, prosodic features improved only modestly over acoustic confidence features alone (less than 7% error reduction).

2.2. Experimental Background

In this section, we provide the basic approach of the confidence scoring module, the speech data, and the labeling of the data. For more details, please see [6, 7, 8].

The confidence scoring module, developed by Hazen *et al.*, is based on a Bayesian formulation. For each recognition hypothesis, a set of confidence measures are computed to form a feature vector, which is reduced to a single dimension using a simple linear discrimination projection. Distributions of this raw confidence score for correct and incorrect hypotheses are obtained from the training data. A probabilistic confidence score is then obtained using maximum *a posteriori* probabil-

mean_F$_0$	mean F_0 of all vowels
max_F$_0$	maximum F_0 of all vowels
mean_pv	mean probability of voicing of all vowels
mean_energy	mean RMS energy of all vowels
max_energy	maximum RMS energy of all vowels
duration	duration of the utterance
pause1_duration	duration of silence before the utterance
pause2_duration	duration of silence after the utterance
%_silence	percentage of silence (as indicated by sum of inter-word pause durations)
speaking_rate	sum of expected vowel durations over sum of measured vowel durations
num_syllables	the number of syllables in the utterance
tempo	number of syllables / total duration

Table 1: Utterance level prosodic features used in experiments

ity (MAP) classification, with the raw confidence score as the input. The threshold of the MAP log likelihood ratio can be varied to set the operating point of the system to a desired location on the *receiver-operator characteristic* (ROC) curve, to balance between high detection rate and low false alarm rate.

Hazen's confidence models used 15 features for detecting utterance-level recognition errors, and 10 features for detecting word-level recognition errors. These features measure how well the input speech fits the underlying models used by the system, as well as whether there are many competing hypotheses that have similar scores. For example, the total utterance score (i.e., the sum of acoustic, language model, and pronunciation model scores) for the top sentence hypothesis measures the overall quality of this hypothesis, while the drop in total score between the top hypothesis and the second hypothesis in the N-best list measures the "distance" between competing hypotheses. The complete inventory of the 25 utterance and word features can be found in [8]. These features, which will henceforth be referred to as ASR (Automatic Speech Recognition) features, are used to train baseline utterance and word confidence models, to be compared with confidence models augmented with prosodic cues. The comparison will be based on the *figure of merit* (FOM), i.e., the area under the ROC curve, and the minimum classification error rate.

In training, an utterance is marked as incorrectly recognized if there are *any* errors in the best sentence hypothesis. This follows the example in Hirschberg's experiments, thus promoting direct comparison. Only words in the top sentence hypothesis are used for training. About 39.4% of the 2334 test utterances had at least one error, and 16.6% of the words were incorrect.

2.3. Utterance-level Experiments

We have selected twelve utterance-level prosodic features as potential candidates for predicting speech recognition errors, as described in Table 1. Three features are related to F_0, two are associated with energy, and the remaining seven features capture various kinds of timing information. F_0 is determined completely automatically for all of our experiments, and is computed using the algorithm described in [9].

There were differences in both the means and variances of the prosodic measurements between correctly and incorrectly recognized user turns, with the variances generally larger for misrecognized utterances. Given that the confidence scoring module uses a probabilistic framework, we believe that a mutual information measure will be a good indication of the effectiveness of each confidence feature.

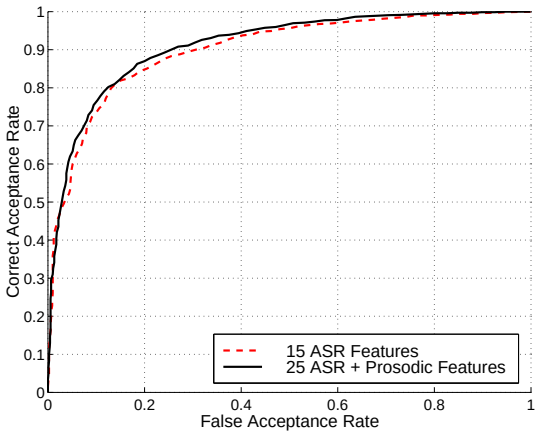

Figure 1: ROC curves of utterance-level speech recognition error detection using only ASR features and using both ASR and prosodic features.

System	FOM	MER	Significance
Baseline	.900	16.9 %	
+ Prosodic Features	.912	15.6 %	*.018*

Table 2: Figure of merit (FOM) and minimum classification error rate (MER) for the utterance-level confidence scoring with only ASR features and with ASR and prosodic features combined. The McNemar significance level between the two classification results is also listed.

We computed the mutual information between each utterance feature and the utterance correctness label, for both the ASR and the prosodic features. The features with the highest mutual information are from the ASR system. This is not surprising, because the ASR features are directly linked to the performance of a recognition system. Nevertheless, some prosodic features also provide significant information about the labels. In particular, the percentage of silence within an utterance, average and maximum F_0 values, utterance duration and tempo are among the "best" prosodic features.

We compared the performance of utterance-level accept/reject decisions with only ASR features and with ASR and prosodic features combined. All fifteen ASR features improved the performance on the development data when added incrementally to the feature set. If both ASR and prosodic features are used, 25 out of 27 features improved the performance on the development data when added incrementally.

Figure 1 plots the ROC curves of the utterance-level classification experiments on the test data. The addition of prosodic features pushed the ROC curve towards the upper-left corner slightly. The figure of merit and the minimum classification error rate are summarized in Table 2 for the two system configurations. The McNemar significance level between the two classification results is 0.018. Thus, the improvement is statistically significant given a 0.05 threshold.

2.4. Word-level Experiments

We have examined nine *word-level* prosodic features as potential candidates for predicting word-level speech recognition errors, which are directly analogous to the utterance-level fea-

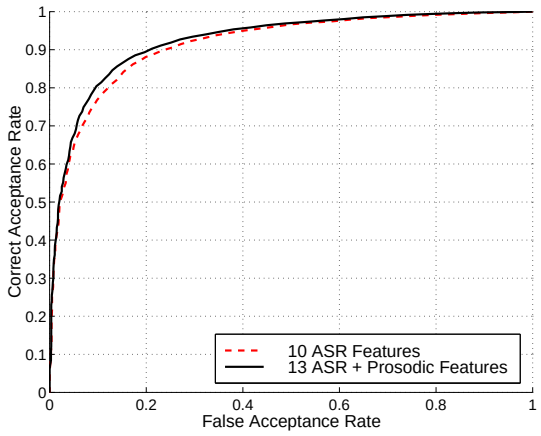

Figure 2: ROC curves of word-level confidence scoring using only ASR features and using both ASR and prosodic features.

System	FOM	MER	Significance
Baseline	.913	10.9%	
+ Prosodic Features	.925	10.2%	*0.0005*

Table 3: Figure of merit (FOM) and minimum classification error rate (MER) for the word-level confidence scoring with only ASR features and with ASR and prosodic features combined.

tures. Three features are related to F_0, two features are related to energy, and the remaining four features capture various timing information of a hypothesized word.

As in the case of the utterance-level features, we computed the mutual information between each word feature and the word correctness ratio. The word energy features, which have been normalized by the maximum utterance energy, are among the "best" prosodic features. This is possibly because they are good indications of background speech, as discussed previously. As for utterance-level features, the top word features are all ASR features. However, prosodic features compare favorably to some ASR features; and more importantly, they provide independent information, and hence, are more likely to bring additional gain.

We obtained the performance of word hypothesis error detection with only ASR features and with ASR and prosodic features combined. All ten ASR features improved the detection performance on the development data when added to the feature set. In the experiment which used both ASR and prosodic features, only the top 13 features improved detection performance on the development data. Seven out of the 13 were prosodic features.

Figure 2 plots the ROC curves of word-level classification experiments on the test data. As shown in the figure, the addition of prosodic features also pushed the ROC curve towards the upper-left corner slightly. Table 3 summarizes the FOM and MER for the two system configurations. The McNemar significance level between the two classification results is 0.0005, which implies that the difference is statistically significant.

3. Phrase F_0 Models

In the previous section, we demonstrated that prosodic features were able to improve both utterance and word level confidence scores. However, we have also found that utterance or word based prosodic measures are usually noisy. In this section, we develop a different framework, in which we model only part of the prosodic space of an utterance, concentrating on phrases that bear important communicative functions. We believe that such an approach is more robust than trying to characterize the intonation of an entire utterance, especially for spontaneous speech. We want to build acoustic models directly for certain linguistic aspects in an utterance, without using prosodic labels as an intermediate layer. In this way, we can avoid the labor-intensive prosodic labeling process as well as the necessity of predicting prosodic labels from linguistic analyses. We can use data-driven methods to derive distinct F_0 patterns/categories for the linguistic components in our modeling framework, which can be regarded as analogous to prosodic labels. We envision that the phrase models can potentially be applied to score the intonation patterns of recognizer hypotheses, which can in turn be used to resort the N-best outputs for improved recognition accuracy or to support the rejection of erroneous hypotheses.

In this section, we examine the feasibility of such a framework by performing a pilot study on characterizing the pitch contours of some selected English phrases in the JUPITER domain. As a starting point, we select five common types of phrases, such as "what is", "tell me", city names, etc., to carry out our study. These phrases also carry important information, so that they are likely to have a significant impact on the system performance. We seek to answer the following questions in our experiments:

(1) Can we identify phrase classes based on the F_0 contour alone?
(2) Does a phrase-level F_0 pattern generalize across similar but not identical utterances?
(3) Does each phrase class have some set of canonical patterns?
(4) Are there interdependencies among phrases in an utterance?
(5) Will this information be useful to speech recognition?

3.1. Related Research

Research on using intonation in the linguistic analysis of spoken utterances has been sparse. Among the few inquiries reported in the literature, most methods rely on an intermediate prosodic transcription to serve as a bridge between the acoustic realization of the intonation and the syntactic/semantic structure of the utterance [10, 11]. These methods need to address several difficult issues. First, prosodic transcription, e.g., using the ToBI convention for English [12], is a challenging and time-consuming task, which makes it impractical to transcribe large speech corpora manually. Secondly, automatic recognition of intonational events (especially pitch accents, phrase tones, etc.) from the acoustic signal is difficult and error-prone [13]. Third, the mapping between prosodic events and the syntax/semantics of an utterance is still poorly understood, except for a general correspondence between prosodic phrase boundaries and syntactic boundaries. For this reason, most studies have focused on using prosodic phrase boundary locations to resolve syntactic ambiguities [10, 14] or to improve parsing efficiency [11]. Although there have been efforts towards automatically describing and classifying intonation contours [15, 16, 13, 17], their use in linguistic analysis or speech recognition has been limited, largely due to the missing link with linguistic identities.

3.2. Experimental Design

One of the key issues in intonation modeling is to find an inventory of model units. In our framework, we want to explore

<**what_is**>:	what is, how is, ...
<**tell_me**>:	tell me, give me, show me, ...
<**weather**>:	weather, forecast, dew point, wind speed, ...
<**SU**>:	Boston, Paris, Monday, ...
<**US**>:	Japan, Detroit, tonight, ...

Table 4: Five common phrase classes and examples for each class in the JUPITER weather domain.

the feasibility of directly modeling certain linguistic structures in English utterances. Thus, we begin with a number of common phrases in the JUPITER utterances. In this way, the unit set covers some "typical" basic linguistic patterns, and there will be sufficient data for acoustic model training.

We have chosen only *two-syllable* phrases in our study, mainly to evaluate the feasibility of our proposed approach. Five classes of two-syllable words/phrases are selected, including "<what_is>", "<tell_me>", "<weather>", "<SU>", and "<US>". Each "phrase" class consists of a list of words/phrases with the same stress pattern, which have also been chosen to have similar semantic properties, so that they are likely to serve similar syntactic functions. In particular, each phrase class consists of words that can be substituted into the following sentence template to produce a well-formed sentence:

<**what_is**> | <**tell_me**> **the** <**weather**> **in**|**for**|**on** <**SU**> | <**US**>

For example, the "<weather>" class contains words or compound words like "weather", "forecast", "wind speed", "dew point", etc., all of which have "stressed unstressed" stress pattern and refer to some kind of weather information; the "<US>" class consists of "unstressed(**U**) stressed(**S**)" two-syllable words for place names or dates; while the "<SU>" class consists of "stressed(**S**) unstressed(**U**)" two-syllable words for place names or dates. Example words/phrases in each class are listed in Table 4.

Utterances that match exactly the above sentence template in the JUPITER corpus are chosen to form a test set. We will conduct experiments to classify the intonation contours of the five phrase classes on this set, and to study the correlation of the intonation contour patterns among the phrases in these utterances. To ensure similarity between the training and test data for the five phrases, an instance of a phrase is used for training only if it occurs at particular positions in an utterance. Specifically, the "<what_is>" and "<tell_me>" phrases are constrained to be from the beginning of an utterance; the "<weather>" phrase is limited to be from an intermediate position in an utterance; and the "<SU>" or "<US>" phrases are selected only from the end of an utterance. Thus, the training set consists of utterances which contain the five phrases at positions described above, excluding those that match exactly the test sentence template. In this way, we can ensure the independence of training and test data and examine if the phrase F_0 patterns can generalize across similar but not syntactically identical utterances. The data selection criteria are summarized and illustrated by some example training and test utterances in Table 5.

To limit the scope of our initial investigation, we use only F_0-based measurements to characterize the phrase intonation pattern. We describe the F_0 contour of a phrase using its constituent syllable F_0 contours, each of which is characterized by the F_0 average and slope. The F_0 contour for each syllable is measured from the sonorant region only, which is determined

Test

what_is|**tell_me** the **weather** in|for|on **SU**|**US**
What is the *weather* in *Detroit*?
Give me the *wind speed* for *Friday*.

Train

what_is|**tell_me** ...
What is the humidity in Honolulu Hawaii?
Give me the *weather* for Chicago for *tomorrow*.
... **weather** ...
Yes, I would like to know the *weather* in *New York*.
Can you tell me the *sun rise* for anchorage?
... **SU**|**US**
Tell me the *wind speed* for concord new Hampshire *today*.
And what is the time in *Frankfurt*?

Table 5: Criteria for selecting training and test utterances ("|" means "or", and "..." means "any words"). Test utterances are selected to match the "test" template. Training utterances are selected to match any of the three "train" templates but not the "test" template.

from time-aligned phonetic and word transcriptions. Thus, each token will be represented by a four-dimensional vector, consisting of the F_0 averages and slopes of the two syllables.

3.3. Results and Discussions

Phrase classification We first perform classification experiments to examine how well phrases can be identified by their F_0 contours. A principal component analysis is first applied on the collection of training vectors to "whiten" the observation space. Mixtures of diagonal Gaussian models are then trained to characterize the distributions of the rotated feature vectors for the five phrase classes. Maximum likelihood (ML) classification is used, because our purpose is to evaluate the ability to identify phrases based on F_0 information alone, without the assistance/interference of *priors*. The F_0 contour of each utterance has been normalized by its mean value, to reduce variances due to speaker pitch differences.

To examine how well the phrase models generalize from training data to test data, we applied the phrase models to classify the phrases in both the training and the test utterances. The five-class classification accuracy is 60.4% on the training data, and 56.4% on the test data. The performance on the *unseen* test data is only slightly worse than that on the training data. We conclude that there exists information in the F_0 contours of the five phrases that can be used to distinguish these phrases. Detailed classification confusions for test data are summarized in Table 6.

Data clustering We performed K-means clustering on the training tokens to identify any canonical F_0 patterns associated with each phrase class. As in the classification experiments, a principle component analysis is applied on the four-dimensional feature vector prior to the clustering, mainly to normalize the variance on each dimension. In order to select the dominant 3 to 4 contour patterns for each phrase pattern, we trained a diagonal Gaussian model for each data cluster, and classified each token according to its preferred cluster group. Tokens that score poorly against all of the Gaussian models are discarded, and clusters with insufficient counts are pruned. Interestingly,

	<what_is>	<tell_me>	<weather>	<SU>	<US>	# Tokens
<what_is>	**45.82%**	19.22%	25.63%	7.94%	1.39%	718
<tell_me>	16.88%	**53.25%**	15.58%	9.09%	5.20%	77
<weather>	4.91%	0.63%	**68.18%**	12.45%	13.83%	795
<SU>	4.46%	3.42%	15.33%	**52.53%**	24.26%	672
<US>	3.25%	1.63%	13.01%	16.26%	**65.85%**	123

Table 6: Classification results for phrases in the test utterances. The reference labels are shown in the first column, the hypothesized labels for the phrases are shown in the first row, and the number of tokens for each phrase class is summarized in the last column. Overall classification accuracy was 56.4%

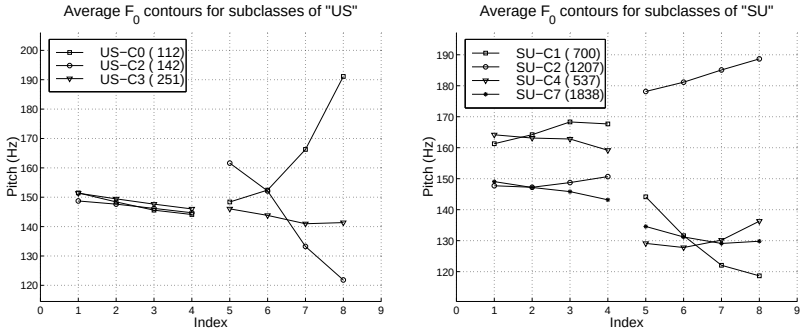

Figure 3: Mean F_0 contours for the "<US>" and "<SU>" phrase clusters, obtained by unsupervised clustering. Each phrase F_0 contour is shown as a sequence of two syllable F_0 contours, each represented by 4 connected samples. The number of tokens in each cluster is given in parentheses after the cluster name.

if we discard from the *training* set all tokens that score poorly against all of the dominant clusters, the 5-class classification performance on the *test* set improves from 56.4% to 58.8%.

Figure 3 shows the dominant cluster groups for the "<SU>" and "<US>" patterns, represented as the mean F_0 contour of the constituent tokens. The subclasses of the "<SU>" and "<US>" phrases are particularly "expressive", possibly due to the fact that these phrases are likely to be accented (because they convey important information such as a place or a date), and they carry a phrase tone as well (because they are at the end of an utterance). There are three patterns for the "<US>" phrase: rising ("US-C0"), falling ("US-C2"), and flat[1] ("US-C3"). It is interesting to note that the contours of the first syllable (unstressed) for the three subclasses are virtually the same, while the large differences among subclasses are only on the second syllable (stressed). This seems to be consistent with intonation theory's view that only stressed syllables are likely to be accented. The first syllable does not carry any intonational events, while the second syllable is responsible for signaling both the accents (if any) and the phrase boundary tone.

The "<SU>" phrase also has the basic rise, fall, and flat patterns. However, the first syllable in the "<SU>" phrase also demonstrates variations, possibly due to its role in carrying pitch accents. In particular, the "SU-C1" and "SU-C4" patterns have higher F_0 levels for the first syllable. We suspect that the first syllable in these two subclasses is more accented. The "SU-C7" pattern is fairly "plain", and its mean F_0 contour is very similar to that of the "US-C3" pattern.

We listened to some utterances labeled with the "SU-C2" or "US-C0" patterns, and generally perceived a rising (question) intonation. These subclasses possibly correspond to the $L^*\ H^-H\%$ and $L^*\ L^-H\%$ patterns described in the ToBI labeling convention [12]. However, we are unable to systematically relate these acoustically derived classes to categories defined in prosodic labeling conventions. It will be interesting to perform the data clustering experiment on prosodically labeled data to facilitate such comparisons.

[1]The slightly falling slope of this subclass is likely due to an overall F_0 declination.

Correlations of phrase patterns We have identified a set of canonical F_0 patterns for each phrase using a data-driven approach. We now use these subclasses to examine any correlations among the acoustic realizations of the phrases within an utterance, e.g., to determine whether certain subclasses of the "<what_is>" phrase are more likely to occur together with certain subclasses of the "<SU>" phrase. The test set is used to carry out this study, because the utterances in the test set are more homogeneous, and each contains exactly three phrases.

We use the *mutual information* measure to quantify the correlation, which is based on the frequency counts of the phrase subclasses in the test utterances. We trained a diagonal Gaussian model using the training tokens in each phrase subclass, and classified the phrases in the test utterances into one of the subclasses. We then tabulated the number of each individual subclass and the number of each subclass *pair* in the test utterances, to compute the mutual information value.

Table 7 shows the results for each pair of "<what_is>" and "<weather>" subclasses, computed using 610 test utterances. Figure 4 shows the contour for the *most compatible* subclass pair, and Figure 5 shows the *most incompatible* subclass pair.

The "weather-C0" subclass (with a high, falling mean F_0 contour) seems to have strong "preferences" with regard to other phrase subclasses. For example, the mutual information between "what_is-C6" (with a very high, rising mean F_0 contour) and "weather-C0" is 0.67. From the mean F_0 contours of "what_is-C6" and "weather-C0" shown in Figure 4, it seems

	weather-C0	weather-C1	weather-C3
what_is-C1	-0.16	0.42	-0.29
what_is-C4	**-0.58**	0.06	0.06
what_is-C6	**0.67**	-0.15	-0.10
what_is-C7	0.12	-0.28	0.12

Table 7: Mutual information between "<what_is>" and "<weather>" subclasses calculated for phrases in utterances matching "<what_is> the <weather> in|for|on <SU>." Mutual information measures larger than 0.5 or smaller than -0.5 are highlighted in **boldface**. A total number of 610 utterances were used in the computation.

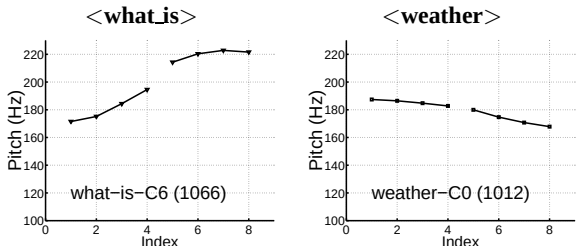

Figure 4: Example of a compatible subclass pair (mutual information = 0.67).

that these two F_0 patterns may form one intonation phrase. On the other hand, the mutual information between "what_is-C4" (with the lowest F_0 among all subclasses of "<what_is>") and "weather-C0" is -0.58, which suggests that these two patterns are highly incompatible. The mean F_0 contours of "what_is-C4" and "weather-C0" are shown in Figure 5. We think that it is difficult (and unnatural) to start a "<weather>" word from an F_0 onset that is higher than the F_0 offset of the preceding syllable.

Although we are unable to derive a formal linguistic explanation for these observations, it is interesting to know that there exist certain correlations among phrases in an utterance. We have developed a framework which is able to quantify these correlations using statistical methods. Such information can potentially be utilized to provide additional constraints in scoring phrase level F_0 patterns.

4. Conclusions

This paper has described two experiments aimed at utilizing prosodic cues to improve spoken conversational systems. The first experiment demonstrates that relatively simple prosodic measures can enhance the performance of a confidence scoring algorithm, at both the word and the utterance levels. The second experiment was a pilot study whose main goal is to establish a procedure for scoring for phrase-level prosodics without time-intensive and error-prone manual labelling. For a selected set of typical phrase patterns in JUPITER, we achieved a 5-class test-set classification performance of 58.8%, using only F_0 information. It remains to be seen whether this research will yield direct benefits in recognition accuracy or in confidence scoring.

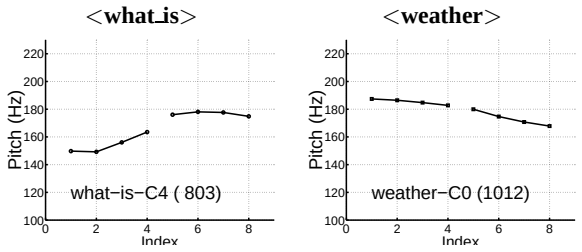

Figure 5: Example of an incompatible subclass pair (mutual information = -0.58).

5. References

[1] Zue, V., S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," in *IEEE Transactions on Speech and Audio Processing 8*(1), 100–112.

[2] Oviatt, S., G.-A. Levow, M. MacEachernn, and K. Kuhn, "Modeling hyperarticulate speech during human-computer error resolution," in *Proc. ICSLP'96*, Philadelphia, USA, pp. 801–804.

[3] Soltau, H. and A. Waibel, "On the influence of hyperarticulated speech on recognition performance," in *Proc. ICSLP'98*, Sydney, Australia.

[4] Hirschberg, J., D. Litman, and M. Swerts, "Prosodic cues to recognition errors," in *Proceedings 1999 IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, USA.

[5] Hirschberg, J., D. Litman, and M. Swerts, "Generalizing prosodic prediction of speech recognition errors," in *Proc. ICSLP'00*, Beijing, China.

[6] Kamppari, S. O. and T. J. Hazen, "Word and phone level acoustic confidence scoring," in *Proc. ICASSP'00*, Istanbul, Turkey, pp. 1799–1802.

[7] Hazen, T. J., T. Burianek, J. Polifroni, and S. Seneff, "Integrating recognition confidence scoring with language understanding and dialogue modeling," in *Proc. ICSLP'00*, Beijing, China.

[8] Hazen, T. J., T. Burianek, J. Polifroni, and S. Seneff, "Recognition confidence scoring for use in speech understanding systems," in *Proceedings 2000 IEEE Workshop on Automatic Speech Recognition and Understanding*, Paris, France.

[9] Wang, C., and Seneff, S., "Improved tone recognition by normalizing for coarticulation and intonation effects," In *Proc. ICSLP'00*, Beijing, China.

[10] Ostendorf, M., C. W. Wightman, and N. M. Veilleux, "Parse scoring with prosodic information: An analysis-by-synthesis approach," in *Computer Speech and Language 7*, 193–210.

[11] Kompe, R., A. Kieβling, H. Niemann, E. Nöth, A. Batliner, S. Schachtl, T. Ruland, and H. U. Block, "Improving parsing of spontaneous speech with the help of prosodic boundaries," in *Proc. ICASSP '97*, Munich, Germany, pp. 811–814.

[12] Silverman, K., M. B. Beckman, J. Pirelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proc. ICSLP'92*, Banff, Canada, pp. 867–870.

[13] Ostendorf, M. and K. Ross, "A multi-level model for recognition of intonation labels," in Y. Sagisaka, N. Campbell, and N. Higuchi (Eds.), *Computing Prosody*, pp. 291–308. Springer.

[14] Hunt, A., "Training prosody-syntax recognition models without prosodic labels," in Y. Sagisaka, N. Campbell, and N. Higuchi (Eds.), *Computing Prosody*, pp. 309–325. Springer.

[15] Grigoriu, A., J. P. Vonwiller, and R. W. King, "An automatic intonation tone contour labelling and classification algorithm," in *Proc. ICASSP'94*, Adelaide, Australia, pp. 181–184.

[16] Jensen, U., R. K. Moore, P. Dalsgaard, and B. Lindberg, "Modelling intonation contours at the phrase level using continuous density hidden Markov models," in *Computer Speech and Language 8*, 247–260.

[17] ten Bosch, L. F. M., "Automatic classification of pitch movements via MLP-based estimation of class probabilities," in *Proc. ICASSP'95*, Detroit, USA, pp. 608–611.