PITCH RANGE MODELLING: LINGUISTIC DIMENSIONS OF VARIATION

David Patterson and D. Robert Ladd
University of Edinburgh

ABSTRACT

This paper reports on a large scale study of pitch range variation across speakers. The experiment examined the relation between a model of pitch range based on pitch level and pitch span with the perception of speaker characteristics. The key to our measure of range is that it is based on clearly defined linguistic targets in speech. These targets included sentence initial peaks, accent peaks, post-accent valleys and sentence final lows. Our data are based on the ratings of 48 listeners judging 32 speakers of British English. The results show that a pitch range model based on linguistic dimensions of variation better captures variation in listeners' judgements than the well established measures based on speakers' long term distributional properties of f0, such as $\pm$ 2sds mean, 95th-5th percentile and 90th-10th percentile.

1. INTRODUCTION

Within psychology and paralinguistics, research on speaker characteristics and on the expression of emotion in speech has explored the communicative effects of pitch range, as well as variation in range both within and across speakers [9]. Despite clear intuitions that pitch range is important in communicating information both about speakers' stable characteristics (e.g. monotonous speakers perceived as boring) and about transient speaker emotions or states (e.g. "raising the voice" to signal anger or surprise), there is little agreement on any but the most coarse- grained findings about the effects of pitch range. It has been suggested elsewhere ([7],[6]) that one of the reasons for this lack of agreement has been the inadequate - and incompatible - definitions of pitch range used in many past studies.

Within linguistics and phonetics, there is similar uncertainty as to what specific properties the term "pitch range" refers to, and there is disagreement aobut how to describe pitch features in tonal and intonational phonology. Ladd [6] has identified two main approaches: the "initialising approach", which describes features as relative to other parts of an utterance, and the "normalising approach", which describes features as relative to the speaker's range. The normalising approach requires a notion of pitch range so that differences between speakers can be abstracted away from, in order to show the invariant properties of the underlying tonal targets. There is evidence supporting the normalising approach (summarised in [6]), but to give full weight to this evidence it is important to have a clearer understanding of pitch range: what it is, how it is measured, and what do modifications of it mean?

For example, pitch range has been defined in [3] as the difference between minimum and maximum f0. This single parameter conveys no information about the distribution of f0 values within that range. Pitch range has also been described [6] using two partially independent dimensions of variation called overall level and span , exemplified in figure 1. From this figure it can be seen that utterance final low would be a reasonable target for measuring level. This target is largely unaffected by raising or lowering of the voice or by other within-speaker range changes. A span measure would ideally encompass the pitch range that is called the tonal space [6]. Although these terms, span and level, have clearly been used with linguistic targets as the key to their identity, measures of f0 level and span based on the long term distributional properties of speech have often been used to characterise voices but with only fairly weak results [5]. The mean and standard deviation do not adequately capture important differences in the pitch range of different speakers [7]. Such measures also do not account for the fact that f0 values often have a non-normal distribution, and also ignore the linguistic motivation for a measure that equates target levels in speech across speakers.

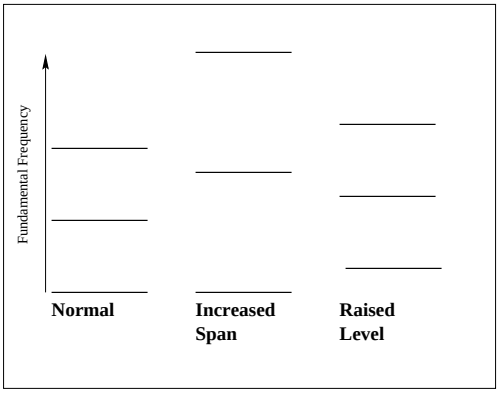

Figure 1: Possible variations in span and level measures

The project summarised here aims to identify an appropriate quantitative characterisation of pitch range, based on mean values of specific linguistic targets, which gives a clearer account of pitch range variation across speakers and more adequately captures the influence of pitch range on judgements made about speakers. A more insightful quantitative model of pitch range will not only be of benefit to the theory of intonation, it also has important consequences in such potential areas as the synthesis of natural sounding speech and automatic speech recognition.

2. METHOD

Our basic approach is to present listeners with recorded speech samples and to correlate judgements of speaker characteristics with a variety of measurable pitch range parameters. Central to our argument, it is assumed that the best measures of level and span are those which show the strongest correlations with the listener judges' data.

One concern with the idea of listeners making judgements on speaker characteristics was that semantic content or social/regional accent would influence the ratings, rather than getting a pure voice judgement. In a preliminary study, we controlled for this by asking

judges to rate voices speaking a foreign language. For this we used a Dutch speech corpus. This corpus was of benefit to the project as pitch range data had already been collected by Shriberg et al [10] using similar points of measurement to those being investigated in the current study. A group of British subjects rated 11 speakers on 20 phonetic and pragmatic criteria such as *confident, tense, irritated, relaxed* and *deep*. The results for listeners' judgements of these phonetic and pragmatic criteria were reduced to 3 dimensions that patterned similarly to those found in previous studies [8]. Importantly for our purposes, ratings on these dimensions correlated better with the pitch range model parameters from the Shriberg et al study than with simple max-min or mean/s.d. measures of pitch range.

The main study is based on listeners' judgements of English speech rated by British listeners. For this experiment careful consideration is necessary to take into account the possible effects of semantic content on listeners' judgements as well as possible positive or negative bias toward regional accent. Consequently, in an attempt to factor out effects of voice quality from "pure" pitch range effects, we asked listeners to judge low-pass filtered speech as well as normally presented speech. If judgements made about voices that are presented in a degraded fashion are similar to those judgements made about voices in a normal presentation, this adds to the evidence that supports the independent function of f0 range in signalling speaker affect.

2.1. Speech materials

Speech data consisted of recordings of 32 speakers, 16 with Standard Scottish accents and 16 with RP accents. Each group contained 8 adult male and female speakers covering a broad range of ages. Nine prose passages, each of about a minute in duration, were read aloud in a single recording session by each speaker. From these recordings, two passages were selected to collect pitch range data for each speaker and to be presented to listener judges. Both passages were considered to be neutral and unemotional. F0 values were extracted at hand-marked locations corresponding to sentence initial peaks (H), other accent peaks (M), valleys (L) and sentence final lows (F) as shown in figure 2. These are well established linguistic landmarks in speech, and impressionistically seemed to be locations that provided fairly stable measures of *different* speaker levels in the speech signal. Further statistical evidence supporting this impressionistic view is given in other related work [2].

Recordings of each speaker were prepared for presentation in one of four different conditions; 2 passages x 2 acoustic conditions (filtered and normal). The four different presentations were prepared on a Latin Square design. The recordings of speakers for each experimental presentation were put in a pseudo random order, the only structure imposed on the order being that in every block of four speakers, each of the four different presentation conditions were represented.

2.2. Rating Form

The choice of adjectives to be investigated was based on those that proved the most reliable in previous similar studies, including our own preliminary Dutch study. The adjectives chosen were *confident, tense, harsh, expressive, deep, weak, irritated, happy, afraid, relaxed, emphatic* and *bored*. For each stimulus, subjects were asked to judge how much each adjective characterised the voice on a seven point unipolar scale from NOT AT ALL to VERY.

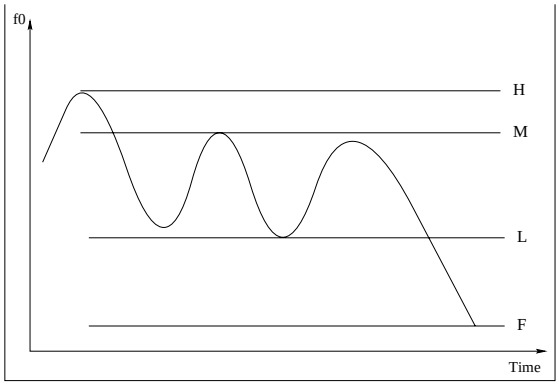

Figure 2: Measurement locations for span and level parameters on an idealised speaker contour

2.3. Subjects

Subjects were of varying ages taken mainly from the student population of Edinburgh University. Each of 4 presentations were judged by 6 English and 6 Scottish listener judges, for a total of 48 subjects.

2.4. Procedure

Listeners were given a practice run based on the speech of two extra speakers who were not part of the main experiment. Listeners were thus well prepared for how much time it would take to fill out a rating form for each speaker, and most importantly they were able to get used to the challenge of making judgements on degraded speech. After the listener judges completed the two practice examples and once the experimenter had made sure that everyone was confident of the task ahead, the full experiment was run without any additional breaks, taking about 45 minutes.

3. RESULTS AND DISCUSSION

From all the pitch data collected at the measurement locations identified in figure 2, various possible parameters corresponding to "level" and "span" were correlated with the listener judgement data. For f0 level, we considered two possible measures of the effective bottom of the range, namely L and F. For f0 span, we considered four possible measures, H-F, H-L, M-F, and M-L. In addition we studied the effect of describing these measures using a linear scale (Hertz), a logarithmic scale (semitones) and the intermediate ERB scale [4]. The correlations between the range data and listener judgement data were carried out on a normalised version of the listener judgement data. The 7 point scale was rescaled on to a new 10 point scale using a simple linear transformation. The normalisation mapped the lowest number that a listener used across all speakers for each individual characteristic to 0 and the highest number that a listener used across all speakers for each individual characteristic to 9. The remaining ratings used between the each listener's minimum and maximum were mapped evenly

	Normal Speech						Filtered Speech					
	Level		Span				Level		Span			
Feature	L	F	M-L	M-F	H-L	H-F	L	F	M-L	M-F	H-L	H-F
afraid	**0.402**	**0.473** √	**−0.603** √	**−0.599**	**−0.524**	**−0.557**	**0.585**	**0.622** √	**−0.676** √	**−0.371**	**−0.616**	**−0.592**
bored	0.141	0.156	**−0.642**	**−0.527**	**−0.652** √	**−0.601**	-0.051	0.033	**−0.785**	**−0.729**	**−0.807**	**−0.830** √
confident	-0.310	**−0.377** √	**0.666** √	**0.641**	**0.639**	**0.655**	-0.321	**−0.362** √	**0.748** √	**0.534**	**0.732**	**0.705**
deep	**−0.926**	**−0.927** √	0.069	0.041	-0.083	-0.083	**−0.967** √	**−0.929**	-0.005	**−0.423** √	-0.092	-0.142
expressive	-0.271	-0.319	**0.754** √	**0.675**	**0.719**	**0.702**	-0.228	-0.299	**0.796** √	**0.627**	**0.752**	**0.762**
emphatic	-0.197	-0.260	**0.799** √	**0.738**	**0.766**	**0.764**	-0.060	-0.126	**0.850** √	**0.729**	**0.821**	**0.818**
harsh	0.277	0.272	0.082	0.079	0.209	0.196	0.046	0.050	-0.277	0.202	-0.295	-0.269

Table 1: Strength of correlation for 2 linguistic measures of level and 4 linguistic measures of span with listener judges' ratings of 7 speaker characteristics for both normal and filtered speech. In this table, as in all the tables in this paper, all correlation coefficients that reach at least a significance level of $p < 0.05$ are in bold. The correlation that is the strongest of the competing measures of level and span for each adjective is marked with a bold tick.

within this 10 point range. This procedure normalised away from variation in the extent that listeners were prepared to use the original scale. For example, listener A used the full rating scale between 1 and 7, while listener B only used 3 through to 6. In these cases the 1 and 3 would both be scaled to 0, while 7 and 6 would be scaled to 9. On this new scale listener A's use of 4 on the original rating scale would be considered different to listener B's use of 4.

Of the 12 adjectives used in the study there were clear examples of adjectives acting in a collinear fashion. Using multiple regression analysis the data was reduced to just 7 adjectives which best accounted for the variation in the data. The adjectives used were *afraid, bored, confident, deep, expressive, emphatic* and *harsh*. Table 1 shows a summary of the strength of correlations of various linguistic measures of pitch span with the 7 adjectives. In table 1, which is based on pitch range measures using the ERB scale, it is clear that sentence final low is the slightly better measure of level. The best measure of span is the difference between average non-initial accent peak and average post-accent valley (M-L). It is clear that for some speaker characteristics, notably *afraid* and *confident*, effects of level and span are partially independent. This supports the hypothesis that two linguistically motivated, partially independent dimensions of variation better characterise the communicative effects of pitch range compared to the single dimension of just max-min f0. This is clearly the case for both normal and filtered speech, which lends support to the claim that there is a genuine independent pitch range effect in the characterisations of speakers.

Table 2 shows a comparison of the different ways of measuring span. It is clear that all three types of measure (linear, ERB and logarithmic) capture differences between speakers' pitch span effectively, but generally speaking the ERB and logarithmic measure show stronger correlations with listener judges' ratings. There is not much difference between these latter two measures, though it the logarithmic measure may be slightly more successful.

It is clear from table 3 that a linguistically motivated measure of pitch range is far more successful in capturing the differences between listener judges' ratings of speakers then any of the widely used span measures that have been used previously on a regular basis. This is shown graphically in figure 3 which plots the average rating for *confident* for each speaker with their M-L span measure on a logarithmic scale (using open circles) and with a measure based on the 90th percentile minus the tenth percentile of all f0 for each speaker (using crosses) [1]. There is a clear positive correlation between width of span and more positivity of judgements of how confident a speaker sounds using the linguistic measure of span. Such a clear correlation is not apparent using the usual statistical measures of range.

	Normal Speech			Filtered Speech		
	Span					
Feature	Hertz	ERB	Semitone	Hertz	ERB	Semitone
afraid	**−0.441**	**−0.603**	**−0.706** √	**−0.463**	**−0.676**	**−0.811** √
bored	**−0.550**	**−0.642** √	**−0.626**	**−0.746**	**−0.785** √	**−0.676**
confident	**0.533**	**0.666**	**0.709** √	**0.623**	**0.748**	**0.755** √
deep	-0.258	0.069	**−0.480** √	-0.327	-0.005	**−0.406** √
expressive	**0.627**	**0.754**	**0.764**√	**0.677**	**0.796**	**0.799** √
emphatic	**0.699**	**0.799** √	**0.771**	**0.789**	**0.850** √	**0.764**
harsh	0.203	0.082	-0.092	-0.257	0.277	-0.251

Table 2: Differences in correlations between the M-L span, using 3 different types of linguistic measure, and listener judges' ratings to 7 adjectives for both normal and filtered speech

	Normal Speech			
Feature	90% Range	80% Range	± 2sds mean	M-L
afraid	-0.127	-0.276	**−0.609**	**−0.706** √
bored	0.030	-0.103	**−0.609**	**−0.626** √
confident	0.096	0.159	**0.578**	**0.709** √
deep	-0.215	0.237	0.294	**0.480** √
expressive	0.115	0.244	**0.604**	**0.764** √
emphatic	0.153	0.256	**0.628**	**0.771** √
harsh	-0.052	-0.046	-0.007	-0.092

Table 3: A comparison of correlations between 3 span measures based on long term general distributional properties and 1 span measure based on linguistic properties

4. CONCLUSIONS

The present study has investigated the best way of measuring pitch range in order to account for pitch range effects such as the signalling of affect. Results show that there is more to a measurement of pitch range than just max-min, or just a simple measure based on the long term distributional properties of f0. This gives further support to a linguistically motivated model of pitch range [10]. Nevertheless we consider that for future research we should attempt to find properties of speakers' long term distributions of f0

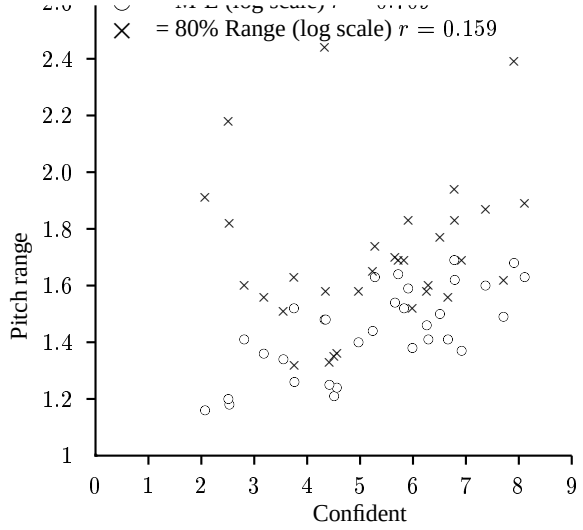

Figure 3: A scattergraph comparing a linguistic measure of span with a regularly used measure of span using general distributional properties of f0 against listener judges' normalised ratings of 32 speakers on the characteristic "confident"

that approximate the parameters used in the model. It will be of great benefit for future research if such properties can be found, as the current method of data collection is very time consuming and labour intensive. As we have shown, such a measure will have to be based on more than just the basic elements of the long term distribution. We will be looking at possible use of skew and kurtosis measures that could well help to capture the range information that we require in our approximation.

ACKNOWLEDGMENTS

This project was supported by a Carnegie Trust for the Universities of Scotland postgraduate scholarship. Many thanks go to Max Louwerse and Alice Turk for their help and advice with the statistical analysis.

REFERENCES

[1] Renee Van Bezooyen. *Characteristics and recognizability of vocal expressions of emotion.* Number 5 in Netherlands Phonetic Archives. Foris, Dordrecht, 1984.

[2] Robert A.J. Clark. Using prosodic structure to improve pitch range variation in text to speech synthesis. In *ICPhS 14*, San Francisco, 1999.

[3] L. Cosmides. Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, 9(6):864–881, 1983.

[4] D.J. Hermes and J.C. van Gestel. The frequency scale of speech intonation. *Journal of the Acoustical Society of America*, 90:97–102, 1991.

[5] W. Jassem. Pitch and compass of the speaking voice. *Journal of the International Phonetics Association*, 1:59–68, 1971.

[6] D. Robert Ladd. *Intonational Phonology*. Cambridge University Press, 1996.

[7] D. Robert Ladd, K.E.A. Silverman, F. Tolkmitt, G. Bergmann, and K.R. Scherer. Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect. *Journal of the Acoustical Society of America*, 78:435–44, 1985.

[8] M. Pakosz. Intonation and attitude. *Lingua*, 56(2):153–178, 1982.

[9] K.R. Scherer. Personality markers in speech. In *Social Markers in Speech*, chapter 5. Cambridge University Press, 1979.

[10] E. Shriberg, D. Robert Ladd, J. Terken, and A. Stolcke. Modeling pitch range variation within and across speakers: predicting fo targets when "speaking up". In *ICSLP*, volume 2, pages 386–389, Stockholm, 1996.