Discrimination of Pitch Change in Speech- and Non-Speech Stimuli

Fred Cummins[1], Colin Doherty[2], and Laura Dilly[3]

[1] Department of Computer Science,
University College Dublin
fred.cummins@ucd.ie
[2] Royal College of Surgeons of Ireland
[3] Massachusetts Institute of Technology

Abstract. This study addresses the interplay between general-purpose auditory processing mechanisms for pitched stimuli and speech-specific processing related to categorical distinctions. We examine the perception of a categorical distinction between certain question/statement pairs that differ physically only in the associated intonation contour in English. Using a standard labeling and discrimination tasks on a continuum ranging from clear statement to clear question, we find a high sensitivity for stimuli around the transition from fall to rise. This does not correspond to the category boundary as revealed by the labeling task. Furthermore, a difference in discrimination sensitivity for speech and non-speech stimuli is found, whereby only the non-speech stimuli can be discriminated reliably when both stimuli are rising, and the previously reported asymmetry, in which stimulus pairs in which the second member has the higher final pitch are discriminated better than the reverse, is found only for non-speech stimuli when both tokens have rising contours.

1 The Processing of Intonation

Speech perception involves the complex interplay of general purpose auditory perceptual mechanisms and speech-specific processing which together allow the recovery of both categorical information and a wealth of gradient, typically non-linguistic, information. The relationship between the two sets of mechanisms remains a central topic of investigation in phonetics, cognitive neuroscience and neurolinguistics.

In this paper, we examine the perception of a categorical distinction between certain question/statement pairs that differ physically only in the associated intonation contour. The pitch contour exhibits a high final rise for questions and a (less steep) fall for statements. This distinction is widely acknowledged to be a clear category distinction in English, German and Dutch, at least. It is of special interest to the investigation of the relationship between general purpose and speech-specific processing, not least because a single physical cue (to a first approximation) underlies the distinction. This contrasts with well-studied consonantal distinctions where a categorical distinction based on manner or place is signaled by a host of cues in parallel [5, 7, 10].

1.1 Categorical Perception and Intonation

Several studies have examined the perception of categorical distinctions cued only[4] by differences in the F_0 contour. Ladd and Morton [6] examined the distinction between "normal" and "emphatic" accent peaks in English. Their methodology was informed by Classical Categorical Perception (CCP) studies in which a continuum is first constructed between clear members of the two categories at issue [4]. Subjects perform an identification (labeling) task for all points in the continuum and a category boundary is inferred from the resulting S-shaped identification function. They then perform discrimination tasks on adjacent pairs from the continuum. The CCP approach is to seek a peak in the discrimination function where adjacent stimuli straddle the inferred category boundary. If this peak is well defined, and discrimination of pairs which do not straddle the boundary is uniformly poor, the case is made that perception is strongly influenced by the categorical judgment being made. The CCP approach is problematic in many ways, and will be discussed further below.

Ladd and Morton found a well-formed S-shaped identification function on their labeling task, but did not find a clear peak in the discrimination function at the inferred category boundary. Furthermore, they observed an interesting and previously undocumented asymmetry in discrimination performance. Stimulus pairs in which the second member had the higher F_0 value (AB pairs) were discriminated with much more success than the reverse, BA, pairs. They replicated this finding in several experiments, but did not offer an explanation, beyond suggesting that there might be a link to the known effect of declination on the perception of the relative prominence of accents, where smaller accents which occur later are perceived to be relatively large compared with earlier accents. They suggested, somewhat reluctantly, that the normal/emphatic distinction might be "categorically interpreted" but not "categorically perceived".

Two studies have adopted the CCP approach to the question/statement distinction which we focus on here. In Schneider and Lintfert [13], listeners did standard identification and discrimination tasks where stimuli were derived from a recording of the sentence "Steht alles im Kochbuch". Identification results confirmed that distinct categories were involved, with individual category switches all lying within 2 steps along the continuum, which corresponded to a difference of less than 30 Hz at the stimulus endpoints. Discrimination results were less clear cut. There was a broad plateau in the middle of the continuum, with poorer discrimination for the more extreme stimuli. An AB/BA difference was also apparent, with worse discrimination for BA, as found also by Ladd and Morton. The link between inferred category boundary and discrimination performance was weak or non-existent, leading the authors to suggest that there might be a third, 'hidden', category between the falling statement and the sharply rising question.

[4] A caveat: Resynthesis with an altered F_0 countour will introduce spectral changes across the board. The claim that the distinction is cued by a single physical variable is thus only true to a first approximation.

In [12], the same approach was taken with Dutch. Again, each subject exhibited a clear categorical response in the identification task, and again the discrimination functions did not support a standard CCP interpretation. In this case, along with the AB/BA asymmetry previously noted, there were two peaks in the discrimination function: one medially, corresponding roughly to the inferred category boundary, and one at the low end of the continuum, corresponding approximately to the point at which stimuli changed from a final fall to a final rise. Despite the author's claim that their results provide a "clear instance of categorical perception of an intonational contrast", no satisfying account of either the low discrimination peak or the AB/BA asymmetry is provided.

The CCP approach has been roundly criticized in [8] for failing to take into account the multiple sources of information that are integrated in order to support speech perception. While this appears to be certainly appropriate for describing the general case of speech perception, the present distinction is of particular interest because a single physical cue (pitch) is implicated in the categorization of a given token as belonging to one class or the other. It must be recognized that other sources of information are, nonetheless, typically available to a listener. Signaling a question using intonation alone, without word order reversal or explicit question marking, is rather unusual, and has a non-neutral pragmatic force suggesting something like incredulity on the part of the speaker. This belief state might well form part of the information used by the listener in decoding a given utterance.

It should be noted in passing that other methods may be employed to establish whether a given hypothetical classification has any basis in reality for listeners. For example, mimicry of continuously graded stimuli may provoke bimodal responses, as in [9, 11]. These may serve to argue that there are underlying linguistic elements which might be uncovered, but they do not inform about the interplay between general auditory processing and speech-specific effects.

1.2 Speech versus Non-Speech

Several brain imaging studies have explored the relationship between the physical cues which signal tone/intonation and brain regions activated. In [2], Gandour and co-workers had English and Thai speakers make discrimination judgments of Thai stimuli which differed in their F_0 contours. The distinction was of linguistic import only for the Thai listeners. Using PET methods, they found many common areas of activation, presumably resulting from common processing of the acoustic stimulus, but only the Thai listeners showed activation of the left frontal operculum, suggesting that this region was selectively engaged when the distinction also reflected distinct linguistic categories.

In similar vein, Gandour et al [3] played speech and non-speech stimuli to Thai and Chinese listeners. The non-speech stimuli were hums matched in pitch to the speech stimuli. Using fMRI techniques, they localized left temporal activation for both groups of subjects in a speech versus non-speech comparison, and in addition, Thai subjects showed additional activation in left inferior prefrontal cortex. This was taken as clear evidence that hemispheric differences in

the perception of speech were sensitive to higher order, domain-specific factors ([3, p. 1082]).

One fMRI study by Doherty et al [1] has also examined the question/statement distinction as signaled by intonation. Three types of stimului were used: rising questions (RQ), Falling Statements (FS) and falling Questions (FQ). The latter required a word order reversal at the start of the sentence, but were otherwise matched to the RQ and FS sentences. Bilateral inferior frontal and temporal activation was found for RQ over FQ and FS, suggesting that these regions are sensitive to the processing of pitch movement, though it is not clear whether this is speech-specific of not. They also found areas in left frontal cortex which appeared to be sensitive to questions (RQ or FQ) but not statements, suggesting that processing of the (categorical, linguistic) distinction between questions and statements was responsible.

1.3 Study Rationale

Given the above results, it was decided to further investigate the nature of discrimination of both speech and non-speech stimuli with rising and falling pitch patterns. Staying within the Germanic language group (English, in this case), we were curious to see whether amassing a greater amount of data would serve to clarify sensitivity to pitch change on both an individual and a group basis. The unexplained asymmetry between AB and BA pairs also seemed to warrant further investigation.

A further motivation for the present study is the question of whether observed discrimination profiles were speech-specific, or whether they would apply to non-speech, pitched stimuli as well. It has been demonstrated that cells at almost all stages of auditory cortex are sensitive to the direction of motion of pitched stimuli [14], and so it seems not unreasonable to assume that a rather general distinction between rising and falling stimuli might be evident for non-speech stimuli too. This would be independent of any putative linguistic categories. This distinction may have been behind the peak in the discrimination function at the lower end of the stimulus continuum found in [12].

2 Methods

2.1 Stimuli

Four types of stimuli were employed, ranging from very speech-like to clearly non-speech. Initially, several repetitions of the first author repeating the isolated word "Norway" with rising, falling and reasonably monotone intonation patterns were made. These served to provide reference values for the endpoints of the stimulus continua used, and one of the monotone recordings was selected as a model utterance for construction of all stimuli used.

For the speech stimuli, the model utterance was resynthesized with an intonation contour which was flat at 100 Hz over the first syllable and then descended

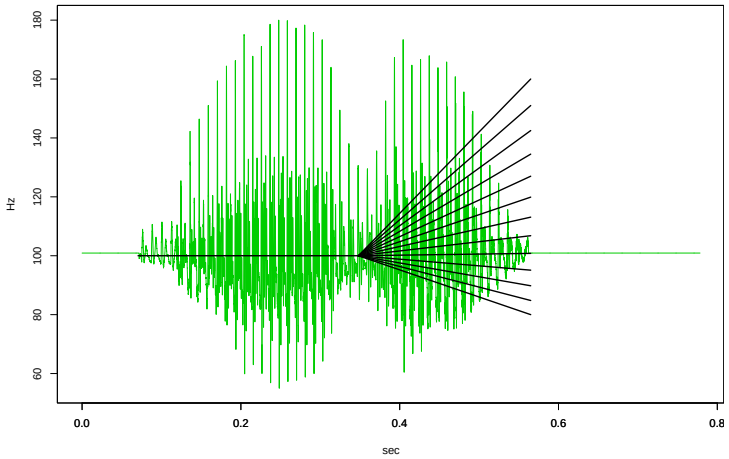

Fig. 1. Pitch contours used. The original speech waveform is shown in the background. Stimuli are numbered from 0 (lowest) to 12. Stimulus 4 is virtually flat.

or ascended linearly to a target value. The lowest target used was 80 Hz and the highest was an octave higher, at 160 Hz. Thirteen distinct points were used, with a one semi-tone difference between consecutive stimulus end points. The stimulus continuum is illustrated in Figure 1. For the most speech-like of the non-speech stimuli, a single pitch pulse was excised from the original speech recording and reproduced many times over. This continuous voiced signal was amplitude modulated to match the original speech token, and the pitch contour resynthesized with the same values as the speech stimuli. This second stimulus set will be referred to as the voiced set.

A third set was made by generating a pulse train which was pitch synchronous with the speech stimuli used, and passing this through a series of linear filters representing five steady-state formants (Praat's 'To hum...' command). The filtered sound was again amplitude shaped to match the speech original. This set will be called the hum stimuli.

Finally, a distinctly non-speech like set was generated in similar fashion, but without the formant-like shaping of the stimuli (Praat's 'To Sound (pulse train)' command). These also had the same pitch pattern and amplitude contour as the speech originals. This final set will be called the buzz stimuli. Samples of all stimuli used, along with Praat scripts to generate them, can be found at http://cspeech.ucd.ie/ ∼fred/intonation/doc/intonationDiscrimination.html.

2.2 Experiment Design

Six native English speakers participated (4f, 2m, ages 22–40). One female speaker was from North-West Canada. All other speakers were from the Republic of Ireland. Results from only three Irish subjects will be presented here. A full report is in preparation. Each subject participated in 4 one-hour trials which took place on distinct days. No subjects reported any known speech or hearing deficit.

On each trial, subjects performed a same/different forced choice task. Where the stimuli were different, they were adjacent stimuli within the 13 point continuum. Stimulus onsets were one second apart, and no repeat hearing was allowed. For each stimulus type, there were 24 possible 'different' trials, in which adjacent stimuli were presented, and these were randomly mixed with 26 'same' trials, giving a basic per-stimulus type block of 50 trials. Sets of four blocks (one per stimulus type) were done consecutively, with a latin square ordering of sets among subjects. Four such sets could be completed in a single hour session, and subjects completed 4 sessions, giving a total of 3200 same/different distinctions per subject. Note that this represents an order of magnitude more data per subject than was used in either [12] or [13]. Stimuli were played through Beyerdynamic DT 100 full cup headphones at a comfortable volume which was constant for all subjects in a quiet, but not sound-treated environment.

At the end of the fourth session, subjects completed an additional labeling trial in which they listened to each of the 13 speech stimuli 6 times in random order and labeled each as being either a question or a statement.

3 Results

In Figure 2, individual response functions are shown for the labeling task. The lower elements of the stimulus continuum are all unambiguously labeled as 'statements', while the high members are labeled 'questions'. The boundary between the two categories lies at approximately stimulus 7 for each subject. The qualitiative boundary at which contours go from falling (stimulus 3), through flat (4) to rising (5) falls clearly within the 'statement' range for each subject.

In Figure 3, the number of correct discrimination responses as a function of stimulus numbers is presented for each of the four kinds of stimulus. Results are pooled across the three subjects. Separate results are shown for AB and BA presentation orders, where 'B' is the higher of the two stimuli in each case. Also shown is the number of false alarms in 'same' trials.

The first result of note is the clear and sharp peak around stimulus number 4 in all cases. This peak clearly marks good discrimination for pairs 3–4 and 4–5. In the stimulus set (see Fig 1), step 4 corresponds almost exactly to a flat pitch contour, so it appears that subjects can discriminate well between the qualitatively different cases of fall vs flat and flat vs rise. There is no AB/BA asymmetry evident for the peak.

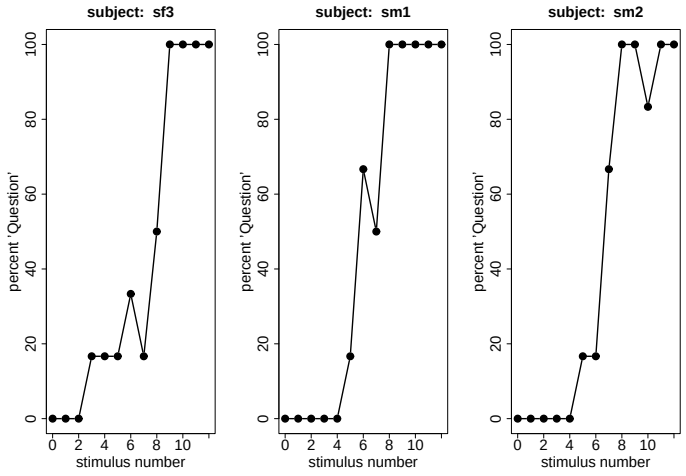

Fig. 2. Percentage of 'question' responses as a function of stimulus numbers for each of three subjects.

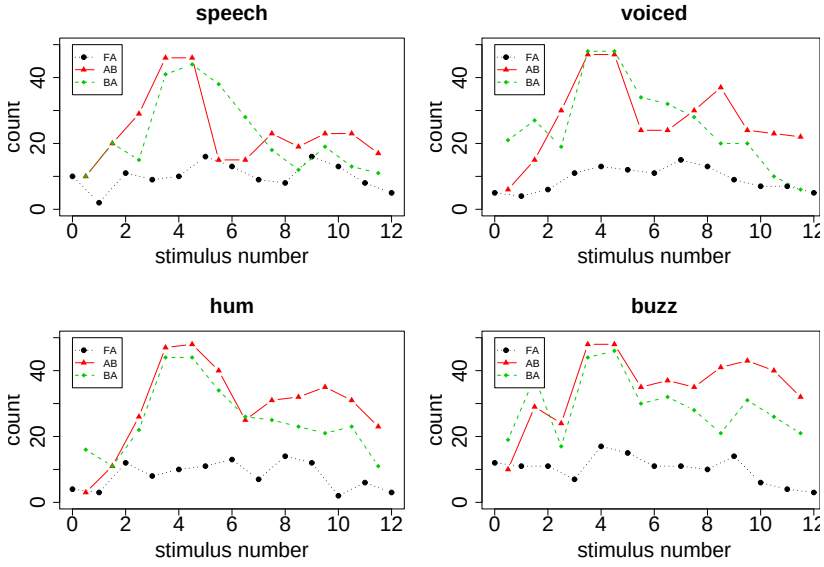

Fig. 3. Number of correctly discriminated 'different' trials for four types of stimulus. Data are pooled from three subjects. In AB/BA discrimination trials, B refers to the higher of the two stimuli. Also show is the number of false alarms (FA).

To the left of the peak, discrimination is poor for all stimulus types, with hits at about the same level as false alarms, suggesting that fall-fall discriminations are hard in all cases. To the right of the peak, however, a different story emerges. Here there seems to be a difference between the speech and non-speech stimuli, which appears as uniformly poor discrimination for rise-rise distinctions with speech stimuli and relatively good rise-rise discrimination with a marked advantage for AB over BA for the buzz stimuli. The voiced and hum stimuli fall in between, corresponding well to our judgement of their relative similarity to speech. False alarm rates appear more or less constant across all positions in the continuum.

4 Discussion

The labeling results obtained support the uncontroversial description of the Question/Statement as a clearly categorical, discrete opposition. It cannot be excluded that intermediate stimuli might represent other, unconsidered, categories, as suggested in passing in [12, 13], but there is no positive evidence to support this claim. The discrimination functions provide no evidence of a third category, as was suggested in the previous studies. No subject mentioned any third interpretation, though that alone may reflect the absence of clear labels for intonational categories, which are often associated with ill-defined pragmatic information. We do not, therefore, further pursue the issue of other potential categories.

The clearest feature of the discrimination functions is the narrow, well-defined peak which straddles the flat element in the continuum (stimulus 4). This is clearly distinct from the apparent boundary between the categories, which lies at about stimulus 7. It seems entirely plausible that this peak arises from a low-level sensitivity to the distinction between rising and falling pitches. This sensitivity is well known from the neurobiology. There is now a strong case to be made that this sensitivity underlies the initial discrimination peak in [12], and that it is explicable without any reference to speech processing as a special case.

In contrast to several previous studies, [6, 12, 13], we did not find a marked asymmmetry in discriminative ability across the board. The asymmetry which appears here is confined to the rising-rising distinctions in the non-speech stimuli. Furthermore, the degree of asymmetry increases as the stimuli become progressively less speech-like. We do not have a full account yet of why this might be. It is not possible, at this remove, to properly gauge the relative naturalness of the stimuli used in previous experiments, all of which had undergone some severe degree of signal processing to obtain the desired pitch contours, but it is conceivable that residual non-speech-like artifacts may account for some of the asymmetry previously observed. Nonetheless, this phenomenon, first reported in [6], is still unaccounted for and is deserving of further study.

Given the observed speech/non-speech differences observed in this study, it seems worthwhile to attempt to localize the speech-specific processing elements involved which appear to inhibit accurate discrimination among rising pairs. One

way to approach this would be to use selected stimuli from the present set in an oddball paradigm ERP study, looking for evidence of selective processing for either the speech or non-speech stimuli.

Acknowledgements

Thanks are due to Sean Connolly and Maresa McGee for discussions around this topic. Work funded by a grant from the Irish Higher Education Authority for collaborative work with Media Lab Europe to the first author.

References

1. C. Doherty, W. C. West, , L. C. Redi, D. Jr Gow, S. Shattuck-Hufnagel, and D. Caplan. The processing of question-intonation: an fMRI study. In *Proceedings of the 15th International Congress of the Phonetic Sciences*, pages 1647–1650, Barcelona, 2003.
2. Jack Gandour, Donald Wong, and Gary Hutchins. Pitch processing in the human brain is influenced by language experience. *NeuroReport*, 9(9):2115–2119, 1998.
3. Jack Gandour, Donald Wong, Mark Lowe, Mario Dzemidzic, Nakarin Satthamnuwong, Yunxia Tong, and Xiaojian Li. A cross-linguistic fMRI study of spectral and temporal cues underlying phonological processing. *Journal of Cognitive Neuroscience*, 14(7):1076–1087, 2002.
4. Stevan Harnad, editor. *Categorical Perception: The Groundwork of Cognition*. Cambridge University Press, 1987.
5. Diane Kewley-Port, David Pisoni, and Michael Studdert-Kennedy. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 73:1779–93, 1983.
6. D. Robert Ladd and Rachel Morton. The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics*, 25:313–342, 1997.
7. Leigh Lisker. *Rabid* vs. *rapid*: a catalogue of cues. *Haskins Laboratories Status Report on Speech Research*, 1985.
8. Dominic W. Massaro. Categorical perception: important phenomenon or lasting myth. In *Proceedings of ICSLP*, pages 2275–2278, 1998.
9. Janet B. Pierrehumbert and Shirley A. Steele. Categories of tonal alignment in English. *Phonetica*, 46:181–196, 1989.
10. Louis C. W. Pols. Variation and interaction in speech. In Joseph Perkell and Dennis H. Klatt, editors, *Invariance and Variability in the Speech Processes*, chapter 7. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.
11. Laura Redi. Categorical effects in production of pitch contours in English. In *Proceedings of the 15th International Congress of the Phonetic Sciences*, pages 2921–2924, Barcelona, 2003.
12. Bert Remijsen and Vincent J. van Heuven. Gradient and categorical pitch dimensions in Dutch: diagnostic test. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 1865–1868, San Francisco, 1999.
13. Katrin Schneider and Britta Lintfert. Categorical perception of boundary tones in German. In *Proceedings of the 15th International Conference of the Phonetic Sciences*, pages 631–634, Barcelona, 2003.
14. Shihab A. Shamma. Auditory cortex. In Michael A. Arbib, editor, *The handbook of brain theory and neural networks*, pages 122–127. MIT Press, 2003.