An Interface for Visualization and Exploration of Spatial Distributions

George M. Shaw
MIT Media Lab
Cambridge, MA 02139
gshaw@media.mit.edu

Deb Roy
MIT Media Lab
Cambridge, MA 02139
dkroy@media.mit.edu

Abstract

This work details methods for visualizing and browsing a set of distributions of locational data derived from multi-modal analysis of a large video and audio dataset.

Introduction

The Human Speechome Project corpus (Roy 2009), (Roy et al. 2006) is a typical large, unstructured dataset. Undertaken with the goal of understanding child language acquisition, the Human Speechome Project (HSP) resulted in approximately 230,000 hours of multi-channel video and audio recorded in a typical home environment over the course of 3 years. This raw data has since been processed, resulting in a corpus of 6,401,748 tokens of transcribed speech and 189,862,525 points representing the tracked locations of people in the video. This processed data accounts for approximately 60% of total recorded audio and near 100% of video. Due to the size of this dataset and its unstructured nature, analysis and exploration requires efficient, scalable methods.

Here we describe methods for tying location data to transcripts, resulting in spatial distributions for various cross sections of speech. We describe visualizations for these distributions and some ways in which these visualizations might be used for analysis. We then describe an interface that has proven useful in working with these distributions based on their statistical properties.

Dataset

One of the current areas of analysis of the HSP corpus is in understanding "spatial grounding," or the ways in which language is tied to particular locations. Language development may be related to the activity contexts in which that language occurs (Bruner 1985), and the locational properties of speech in the home are a useful proxy for this context.

In order to analyze spatial grounding, a set of spatial distributions is derived as follows. 2D histograms are initialized where the bins correspond to discrete locations in the home. Bins are square with sizes ranging from 100mm to 1000mm on a side, and are distributed uniformly throughout the space. For 100mm bin sizes, distributions contain 162 x 118 = 19,116 bins. The 1000mm distributions contain 16 x 11 = 176 bins.

For each utterance of interest, the set of points corresponding to the location of people at the time of that utterance are added to the appropriate bin(s) of the histogram using bilinear interpolation. Each bin is given a weight corresponding to the amount an artificial bin centered at the point would overlap with the bin in question. A weighted point is then added to each bin. Note that by this method, at most 4 bins can be affected by a single point and a point that falls directly in the center of a bin affects only that bin.

Spatial histograms can be thought of as multinomials with the added property that bins have spatial adjacencies, where k = the number of bins and n = the number of samples (in this case utterance points). The probability of an utterance occurring at a location i is the total count of points in $i = X_i$ divided by the total number of points n: $p_i = \frac{X_i}{n}$ and $\sum_i^k p_i = 1$.

The histograms described above were derived for all utterances in the home, and then were further segmented in a variety of ways. In particular, utterances containing 658 "target words" were derived, corresponding to the words in the child's vocabulary at age two. These target word distributions were then further segmented by utterances occurring before the child's first production of a given word, and those after. Finally, all histograms were segmented into those representing each person of interest in the home (including the mother, father, nanny, and child). These histograms taken together form a hierarchical taxonomy where one can examine spatial properties of speech at a variety of levels of granularity and along various dimensions of interest.

Visualization

Several visualization types were initially created from the spatial distributions. At the core are heat maps utilizing a "rainbow" spectrum of color to represent the bin probabilities p_i. These heat maps are normalized such that $max(p_i)$ is depicted in white and $min(p_i)$ is black. These basic heat maps are also extended to heat maps plotted on a log scale, similarly normalized. The log scale versions are useful for displaying more subtlety in cases where there are many points and the range between minimum and maximum counts is large.

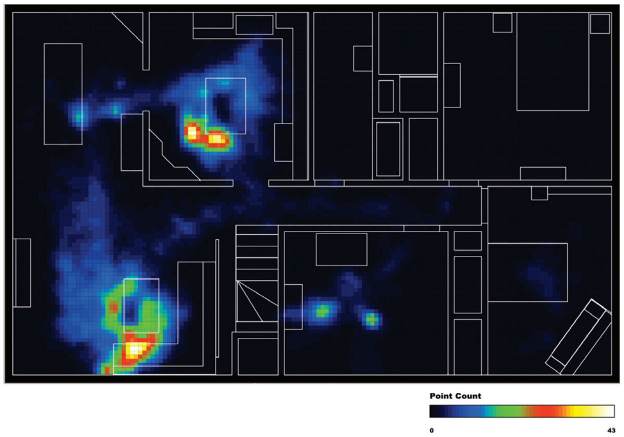

Figure 1: Heatmap for the word "car"

Finally, difference maps are produced that visually represent a distribution's difference from the background. These maps are derived by subtracting the likelihood of each bin in the background distribution P from each bin in the candidate distribution Q:
$D(P, Q) = q_i - p_i \quad \forall i$

Results might therefore be negative, with positive numbers reflecting bins where the candidate distribution is more likely than the background. A modified color spectrum is used in these difference maps where zero is still depicted in black, but positive numbers utilize the warmer end of a rainbow spectrum (red, orange, yellow, and white) and negative numbers are depicted in cooler colors (blue, green). This formulation creates an intuitive depiction, while allowing quick differentiation between positive and negative values.

Analysis

There are many ways in which a researcher might select a set of words to focus on from the larger set of 658 target words. For example, one might like to compare the ways in which noun usage differs from verb usage, or might instead prefer to focus on words that the child learned first. In the current work, the goal was to find words whose usage was focused in one or more discrete locations.

Given a set of interesting words, the next step in analysis is to find words that are similar (or contrasting) to the initial set in their spatial properties. To that end, we might examine these words' spatial distribution heat maps in order to look for clues as to their relationship. As an example, a word that became a focus because of related research was "car." The usage pattern for "car" differs significantly from the overall speech pattern in the home; a difference that shows up immediately in the difference map depicted in Figure 2 - the area near the window in the family room appears as bright yellow and orange, with the rest of the house being primarily blue, black, and green. The highly localized usage pattern evident in the difference map led us to search visually for other similar patterns, revealing words that were localized in similar ways (but in different areas of the home) to "car."

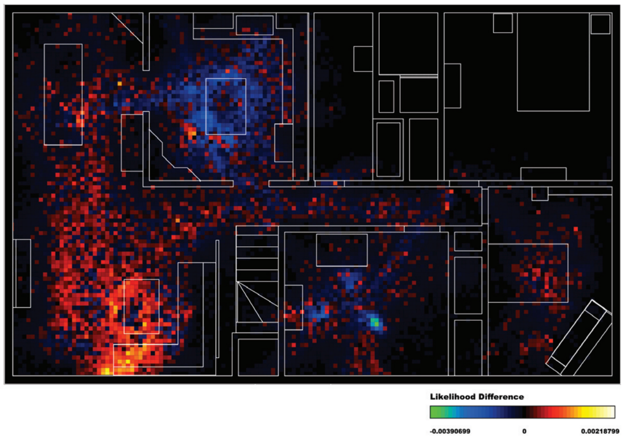

Figure 2: Difference map for "car"

Distribution Browser

Given the insights from visual analysis, the next task is to find statistical measures that describe these differences. A multitude of applicable measures are available, as well as a huge variety of techniques and variations for applying each measure. A few examples that have been explored are entropy, KL-divergence, and Bhattacharyya distance, with the latter two being measured against the background. Spatial measures including Ripley's K (Dixon 2002) and Moran's I (Moran 1950) have been explored as well. Because of the size of the search space for useful measures and the complexity of the dataset, strictly numerical analysis can prove cumbersome and unintuitive. Visualization provides a useful alternative approach.

The localized usage patterns described above can be seen in the difference maps even in very low resolution images, implying that these sorts of differences could be drawn out by looking at aggregate views of all distributions where each distribution is rendered at a small size. As a result of this observation, an approach was devised as follows. All spatial distributions are visualized as small, iconic heat maps. These icons are arranged from top left to bottom right according to a user-specified criteria (i.e. alphabetical). We then apply a statistical measure (i.e. KL-divergence) to each distribution, generating a score for each. Icons are then darkened according to this score. The user can choose to visualize the scores in ascending (low scores are brighter) or descending (high scores are brighter) order. Additionally, the user can choose to filter the distributions by this score, showing, for example, only the top 50 scoring distributions.

Users can switch seamlessly between the various statistical measures, the ordering direction (ascending or descending), and the amount of filtering. The user can also choose to more closely examine any individual distribution in standard, log, or difference form. Additionally, an ordered list is provided for each measure that shows a total ordering of the target words based on the currently selected measure.

One can quickly get a sense of the shape of the distribution over the measure being examined. For measures that provide good separation between spatial distributions, the user sees a uniform spread between dark and light icons. For a measure that clusters distributions toward one end of the scale, however, the user will see an even distribution in the dark (or light) part of the range, and just a few icons at the other end of the range.

As an example of the above effect, a particular measure gives a numerical score to "car" of .90. The next word, "diaper" scores .68. There are 15 words scoring between .02 and .50, and 408 words between 0 and .02. It is clear that most words have low scores, some have higher scores, and "car" is an outlier at the top of the scale. These properties are apparent when viewing the browser, as sorting in ascending orders shows nearly all icons as very bright, with just a few appearing dark, and "car" being black. Sorting in descending order is equally informative, as "car" appears very bright, several icons are less bright, and most icons are dark or black.

The browser allows the researcher to make informed decisions about the best statistical measure to use in order to select desired distributions. In the example of "car," we were able to cycle through many measures quickly, noting in each case the position of "car" along the continuum from dark to light. We were similarly able to look for measures that highlighted words with similar spatial properties (in this case, words whose difference maps appeared tightly clustered in a particular location). As a result, we were able to conclude that the Ripley's K statistic selects the desired spatial distributions. We could then use this measure to automatically sort the 658 target words, as well as any of the 26,000 other words in the corpus' vocabulary.

Additional benefits are realized when we consider the ordering of the icons as a second dimension by which to view distributions, with darkening and lightening as the first dimension. An important current area of interest is in finding spatial information that is predictive of the age at which the child first acquires a given word. To that end, we seek measures that are correlated with age of acquisition. In order to perform this search, we first order the distributions by age of acquisition, and then apply some measure. If correlation is high, we expect to see a smooth transition from dark (or light) at the top left to light (or dark) at the bottom right. Such a transition implies that measure values are varying with age of acquisition. Figure 3 shows such an ordering for 120 words, with KL-divergence applied. We can see that KL-divergence values tend to be lower at the top left (distributions are darker) and higher at the bottom right. Although correlation is not perfect ($r = 0.58$), we can get a quick sense of the appropriateness of the measure. We can also quickly find outliers, or those distributions that are poorly predicted by looking for discontinuities in shading. For example, notice that "round" is far brighter than would be appropriate given its position in the matrix.

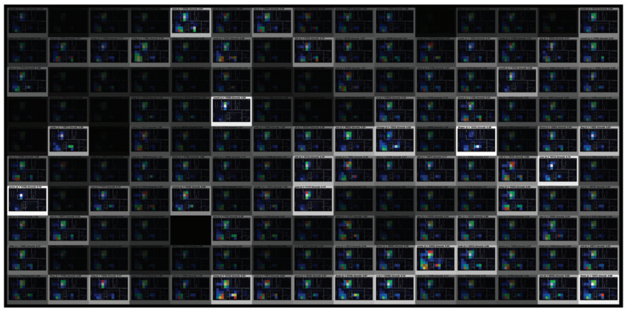

Figure 3: Distribution Browser

Conclusion

This work exemplifies a methodology that seamlessly blends the strengths of computation with those of the human operator, with visualization as the primary catalyst. We produce a set of abstract data structures (spatial distributions) by distilling a large multi-modal dataset (HSP tracks + transcription). We then choose a focus from these data structures based on research goals. By examining the visualizations related to this focus (both those that exemplify it and those that run counter to it) we can begin to discern patterns. We then shift back to looking at the full set of distributions via a distribution browser, attempting to find statistical properties that create an ordering of the distributions that align with the research focus. These properties can then be used at scale in order to generate new foci as well as to provide data points for further research efforts.

References

Bruner, J. 1985. The role of interaction formats in language acquisition. In Forgas, J., ed., *Language and social situations*. Springer. 31–46.

Dixon, P. 2002. Ripley's k function. *Encyclopedia of Environmetrics*.

Moran, P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37(1/2):17–23.

Roy, D.; Patel, R.; DeCamp, P.; Kubat, R.; Fleischman, M.; Roy, B.; Mavridis, N.; Tellex, S.; Salata, A.; Guiness, J.; Levit, M.; and Gorniak, P. 2006. The human speechome project. *The 28th Annual Conference of the Cognitive Science Society* 1–6.

Roy, D. 2009. New horizons in the study of child language acquisition. *Proceedings of Interspeech 2009*.