A thorough evaluation of the Language Environment Analysis (LENATM) system

many[1]

[1]

Author Note

Correspondence concerning this article should be addressed to many, . E-mail:

Abstract

waiting

A thorough evaluation of the Language Environment Analysis (LENATM) system

Brief introduction to LENA(R) products

Previous validation work

Present work

Methods

Corpora

Processing

LENA classification accuracy

Speech and talker segmentation metrics.

Precision and recall.

CVC and CTC evaluation

AWC evaluation

Results

Before starting, we provide some general observations based on the human annotation. Silence is extremely common, constituting 79% of the frames. In fact, 34% of clips contained no speech by any of the human speaker types (according to the human annotators). As for speakers, female adults make up 11% of the frames, the child contributes to 4% of the frames, whereas male adult voices, other child voices, and electronic voices are found in only 1% of the frames each. Overlap makes up the remaining 3% of the frames. The following consequences ensue: if frame-based accuracy is sought, a system that classifies every frame as silence would be 79% correct. This is of course not what we want, but it indicates that systems adapted to this kind of speech should tend to have low "false alarm" rates, i.e. a preference for being very conservative as to when there is speech. If the system does say there is speech, then it had better say that this speech comes from female adults, who provide a great majority of the speech. In second place, it should be key child. Given that male adults and other children are rare, a system that makes a lot of mistakes in these categories may still have a good global performance, because these categories are extremely rare.

LENA classification accuracy: False alarms, misses, confusion

Our first analysis is based on standard speech technology metrics, which put errors in the perspective of how much speech there is. That is, if 10 frames are wrong in a file where there are 100 frames with speech, this is a much smaller problem than if 10 frames are wrong in a file where there is 1 frame with speech. In other words, these metrics should be considered relative error metrics. One problem, however, emerges when there is no speech whatsoever in a given file. In the speech technology literature, this is never discussed, because most researchers working on this are basing their analyses on files that have been selected to contain speech (e.g., recorded in a meeting, or during a phone conversation). We still wanted to take into account clips with no speech inside because it is key for our research goals: We need systems that can deal well with long stretches of silence, because we want to measure how much speech children hear. Indeed, as mentioned above, 38% of our clips had no speech whatsoever. In these cases, the false alarm and confusion metrics are undefined. It also occurred that there was just a little speech; in this case, the denominator is very small, and therefore the ratio for these two metrics ended up being a very large number. Since the presence of outliers violate a basic assumption of regression models, and outliers greatly impact means, we declared as NA any metric that was 2 SD above the mean over all clips. Please note that this leads to an overestimation of LENA's performance, because clips where the relative error rate is very high are removed from consideration. Also, preliminary analyses revealed that performance was lower when near and far were collapsed together (i.e., CHN and CHF were mapped onto a single CH category), so the following analyses use only near speaker categories (i.e., CHN, FAN, MAN, CXN) as well as the overlap category (OLN), with all other categories mapped as non-speech (i.e., CHF, CXF, FAF, MAF, NOF, NON, OLF, TVF, SIL). For a first analysis on all files, TVN was also mapped as non-speech; a follow-up analysis only on ACLEW data segregated TVN such that there were 5 "speaker" categories: CHN, FAN, MAN, CXN, and TVN.

LENA's false alarm (i.e., saying that someone was speaking when they were not) averaged 15%, whereas the miss rate averaged 55%. Imagining for a moment this is a two class solution (speech versus non-speech), then this means that the LENA system avoids "fantasizing" speech that is not there, at the risk of missing speech that is there. This kind of setting is preferable when prioritizing precision over recall. We'll return to that below, when actually discussing precision and recall of the different tags. The confusion rate, as mentioned above, is only calculated for the correctly detected speech (i.e., not the speech that was missed, which counts towards the miss rate, nor the speech that was falsely identified, which is considered in the false alarm). The confusion rate was very low, averaging 8%. These three metrics can be added together into a single "diarization error rate"; of course, if one of them is NA, then DER is NA; 36% of the clips had NA diarization error rate (mostly due to false alarm rate being NA). The mean diarization error rate over all other clips was 79%. In a secondary analysis only on the ACLEW data, … COMPLETE… not sure the evaluation would be fair to LENA. My understanding is that their human annotators marked all sound as TV – whereas you only marked speech as electronic. This means that neither the recall nor the precision can be trusted in our analysis: If we find that 50% of what LENA called TV was tagged as electronic speech, this may well be true. the other 50% was music, jingles, other TV sound. If we say that the recall is 30%, we don't know what the LENA-defined recall was – perhaps LENA did miss 70% of what you tagged as electronic speech, but found 100% of the music and the other TV sounds, so the recall might be much higher than what we say.

LENA classification accuracy: Precision and recall

By now, we have established that the best performance (when "far" labels such as CHF and

Therefore, this subsection shows confusion matrices, containing information on precision and recall, for each key category. For this analysis, we collapsed over all human annotations that contained overlap between two speakers into a category called “overlap”. Please remember that this category is not defined the same way as the LENA overlap category. For LENA, overlap between any two categories falls within overlap – i.e., CHN+TV would be counted towards overlap; whereas for us, only overlap between two talker categories (e.g., key child and female adult) counts as overlap. (Note that neither case contemplates overlap between two speakers of the same category as overlap.)

We start by explaining how to interpret one cell in Figure (precision): Focus on the crossing of the human category FEM and the LENA category FAN; when LENA tags a given frame as FAN, this corresponds to a frame tagged as being a female adult by the human 59% of the time. This category, as mentioned above, is the most common speaker category in the audio, so that over 65k frames were tagged as being female adult by the human and LENA. The remaining 41% of frames were actually other categories according to our human coders: 36% were silence and 5% were confusion with other speaker tags. Inspection of the rest of the confusion matrix shows that, other than silence, this is the most precise LENA tag. Precision for CHN is second, at 40%; thus, fewer than half of the frames labeled as being the key child are, in fact, the key child. The majority of the framesthe LENA incorrectly tagged as being the key child are actually silence (or rather, lack of speech) according to the human annotator (44%), with the 16% remaining errors being due to confusion with other categories: About 9% of them are actually a female adult; 2% are another child; and 5% are regions of overlap across speakers, according to our human coders. MAN and CXN score similarly, 8 and 7% respectively, meaning that less than a tenth of the areas LENA tagged as being these speakers actually correspond to them. As with the key child, most errors are due to LENA tagging silent frames as these categories. However, in this case confusion with other speaker tags is far from negligible. In fact, the most common speaker tag in the human annotation among the regions that LENA tagged as being MAN were actually female adult speech (34%); and, for CXN, it was the key child. In a nutshell, this suggests extreme caution before undertaking any analyses that rely on the precision of MAN and CXN, since most of what is being tagged as such is silence or other speakers. Another observation is that the "far" tags of the speaker categories do tend to more frequently correspond to what humans tagged as silence (77%) than the "near" tags (49%), and thus it is reasonable to exclude them from consideration. The relatively high proportion of near LENA tags that correspond to regions that humans labeled as silence could be partially due to the fact that the LENA system, in order to process a daylong recording quickly, does not make judgments on small frames independently, but rather imposes a minimum duration for all speaker categories, padding with silence in order to achieve it. Thus, any key child utterance that is shorter than .6 secs will contain as much silence as needed to achieve this minimum (and more for the other talker categories). Our system of annotation, whereby human annotators had no access whatsoever to the LENA tags, puts us in an ideal situation to assess the impact of this design decision, because any annotation that starts from the LENA segmentation should bias the human annotator to ignore such short interstitial silences to a greater extent than if they have no access to their tags whatsoever. These analyses shed light on the extent to which we can trust the LENA tags to contain what the name indicates. We now move on to recall, which indicates a complementary perspective: how much of the original annotations were captured by LENA.

```
Again, we start with an example to facilitate the interpretation of this figure: The bes
Many colleagues also use the LENA as a first pass to capture female adult speech via the
For the remaining two near speaker labels (MAN, CXN), recall was 15-18%, meaning that le
```

Finally, as with precision, the "far" categories show worse performance than the "near" ones. It is always the case that a higher percentage of frames is "captured" by the near rather than the far labels. For instance, out of all frames attributed to the key child by the human annotator, 43% were picked up by the LENA CHN label and 0% by the LENA CHF label. This result can be used to argue why, when sampling LENA daylong files using the LENA software, users need not take into account the “F” categories.

Child Vocalization Counts (CVC) accuracy

Given the inaccuracy of far LENA tags, and in order to follow the LENA system procedure, we only counted vocalizations attributed to CHN and ignored those attributed to CHF. As shown in Figure (CVC), there is a strong association between clip-level counts estimated via the LENA system and those found in the human annotations: the Pearson correlation between the two was .7 when all clips were taken into account, and .77 when only clips with some child speech (i.e., excluding clips with 0 counts in both LENA and human annotations) were considered. This suggests that the LENA system captures well differences in terms of number of child vocalizations across clips.

However, users need more: They also interpret the absolute number of vocalizations found by LENA. Therefore, it is important to also bear in mind the relative error rate: Given a LENA estimate, how close may the actual number be? One issue is, as discussed above for the speech technology metrics, relative error rates require the number in the denominator to be non-null. For this analysis, thus, we removed the 306 clips in which the human annotator said there were no child vocalizations whatsoever. When we do this, the mean relative error rate is -37% (median -52%), indicating that the LENA underestimates the number of vocalizations by about a third. However, the range was considerable, going from -100% to 700%. A reanalysis of absolute error rates shows quite a different pattern: FILL THIS IN

Conversational Turn Counts (CTC) accuracy

```
Again, we only considered "near" speaker categories in the turn count, and applied the s
```

A reanalysis of absolute error rates shows quite a different pattern: FILL THIS IN

Adult Word Counts accuracy

One child in the SOD corpus was learning French. We have included this child to increase power, but results without this one child are nearly identical. The association between clip-level LENA and human AWC was strong: Pearson r over all clips was r=.75. Excluding clips where both the human annotators and the LENA reported word counts of zero led to a slightly lower estimate, at r=.69. Mean relative error rates excluding the 361 clips where the human annotators said there were no words (because RER is undefined in such clips) averaged 55% (median -18%), with a considerable range (-100 to 7400%).

Effects of age and differences across corpora

The preceding sections include results that are wholesale, over all corpora. However, we have reason to believe that performance could be higher for the corpora collected in North America (BER, WAR, SOD) than those collected in other English-speaking countries (ROW) or non-English speaking populations (TSI). Additionally, our age ranges are wide, and in the case of TSI children, some of the children are older than the oldest children in the LENA training set. To assess whether accuracy varies as a function of corpora and child age, we fit mixed models as follows. We predicted false alarm, miss, and confusion rates from corpus, child age, and the interaction as fixed, child ID as random, on clips where there was some speech according to the human annotator; since misses and confusions are undefined when the annotator said there was no speech but false alarms is not, we additionally fitted a model to all clips (i.e., regardless of whether they had speech or not. We followed up with an Analysis of Variance (type 3) to assess significance. In none of these analyses was corpus, child age, or their interaction significant. For CVC, we fit a mixed model where CVC according to the human was predicted from CVC according to LENA, in interaction with corpus and age, as fixed factors; with child ID as random effect. An Analysis of Variance (type 3) found a triple interaction, suggesting that the predicted value of LENA with respect to human CVC depended on both the corpus and the child age; and a two-way interaction between CVC by LENA and corpus. To investigate these further, we fit the same regressions within each corpus separately. This revealed that there accuracy of LENA CVC increased with age for BER, but decreased for WAR, being stable in the others.

For CTC, we fit a mixed model where CTC according to the human was predicted from CTC according to LENA, in interaction with corpus and age, as fixed factors; with child ID as random effect. An Analysis of Variance (type 3) found a two-way interaction between CTC by LENA and corpus. To investigate this further, we fit the same regressions within each corpus separately. These follow-up analyses revealed that CTC by LENA was a better predictor of human-tagged CTC for WAR (t=5.06) than ROW (t=2.04) or SOD (t=2.92), and for these than for BER (t=1.07).

Discussion

Acknowledgments

References