Joseph A. Stanley*, Lisa Morgan Johnson and Earl Kjar Brown

Testing the effect of speech separation on vowel formant estimates

https://doi.org/10.1515/lingvan-2024-0152
Received July 26, 2024; accepted December 6, 2024; published online January 30, 2025

Abstract: While recent advances in sociophonetic data processing have made it possible to analyze large datasets and audio not originally intended for linguistic analysis, overlapping speech in recordings with multiple speakers continues to be an issue that results in lost data. We evaluate whether current source separation models produce audio that is clean enough to produce reliable measurements for sociophonetic analysis. We compare formant estimates from a pair of pristine recordings and merged-and-separated versions of those same recordings using the Libri2mix, Whamr16K, and WSJ02mix source separation models. Based on auditory inspection of the separated files, visualization of vowel formant estimates, and statistical analysis, Libri2 performed best and WSJ02 was worst. While the mean formant measurements per vowel were usually small, differences for each observation were larger in unpredictable ways. We are cautiously optimistic about using these tools in sociophonetic analysis, so long as analysis is conducted on vowel means. We conclude with recommendations that researchers can implement when using source separation in sociophonetic research.

Keywords: source separation; sociophonetics; methods; data processing; vowel formant analysis

1 Introduction

Recent advances in sociophonetic data processing have made analyzing large datasets possible. Forced-aligners like MAUS (Schiel 1999), Prosodylab (Gorman et al. 2011), and MFA (McAuliffe et al. 2017) take utterance-level transcriptions and produce word- and phoneme-level alignments. From there, tools like FAVE (Rosenfelder et al. 2014) and Fast Track (Barreda 2021) can automatically extract formant estimates and other acoustic measurements. Transcription has always been a major bottleneck, but tools like Transcriber (Boudahmane et al. 1998) and ELAN (Brugman and Russel 2004) have facilitated manual work, while software like DARLA (Reddy and Stanford 2015) and Bed Word (Ma et al. 2024) and recently developed AI tools (Radford et al. 2023) automatically transcribe audio. Some online interfaces like DARLA (Reddy and Stanford 2015) and BAS Web Services (Kisler et al. 2017) combine several processing steps, allowing for semi- or fully automatic processing.

In addition to speeding up the processing time for traditional sociolinguistic interviews, these advances have facilitated the analysis of audio not originally gathered for linguistic analysis, like public speeches (Bowie 2003; Harrington et al. 2000; Holliday 2024; Wolfram et al. 2016) or personal vlogs (Cheng 2018, 2023; Lee 2017; Mendoza-Denton 2011). In particular, linguists are increasingly turning towards oral histories collected by folklorists and historians to examine how language was used by those born several generations ago (Hickey 2017; Strelluf and Gordon 2024, among many others).

The problem with such recordings is the amount of overlapping speech. Even in sociolinguistic interviews, fieldworkers may inadvertently overlap with their interviewees. One solution is to simply exclude from analysis any speech that overlaps with another person's speech (Olsen et al. 2017). Unfortunately, it takes time to manually code the overlapping speech to try to salvage nearby nonoverlapping speech and it decreases the amount of analyzable data in an already limited dataset.

***Corresponding author: Joseph A. Stanley**, Brigham Young University, Provo, USA, E-mail: joey_stanley@byu.edu. https://orcid.org/0000-0002-9185-0048
Lisa Morgan Johnson and Earl Kjar Brown, Brigham Young University, Provo, USA. https://orcid.org/0000-0002-6817-7713 (L.M. Johnson). https://orcid.org/0000-0002-1121-4931 (E.K. Brown)

However, recent AI tools in speaker diarization and audio source separation appear promising. Speaker diarization is the process by which speaker labels are applied to segments in a single audio track, essentially answering the question "who spoke when?" It is accomplished by extracting speaker embedding vectors based on voice characteristics in the audio file, much like word embedding vectors place words in a multidimensional space such that similar words (e.g., *dog*, *cat*) are close to each other in comparison to words that are different from each other (e.g., *dog*, *flower*; Mikolov et al. 2013). After the speaker embedding vectors are created, they are then grouped into distinct clusters, and then finally, the predicted speaker label is applied to all the segments within a given cluster. This is a first step to source separation, which produces separate audio files with the segments previously identified as belonging to the same speakers, one speaker per audio file (e.g., a WAV file). In theory, good models cleanly separate speakers out from an audio recording with overlapping speech, recovering audio that would otherwise be excluded from sociophonetic analysis. In cases where the amount of data is very limited, such as in archival recordings, any amount of recovered audio will increase the sample size and bolster the claims made from such recordings.

In this paper, we evaluate whether three open-source source separation models produce reliable enough audio for sociophonetic analysis. We do this by presenting results of a case study of vowel formants extracted after applying three source separation models to artificially overlapping speech. This paper therefore serves as a "proof of concept" to show how source separation can fit into the pipeline of sociophonetic data processing.

2 Methods

To test the effect of source separation on acoustic measurements, we took two independent recordings and processed them using typical sociophonetic methods. We then artificially overlapped them into a single track, separated them with three source separation models, and processed those separated versions using the same methods. We used synthetic overlap rather than naturally occurring overlapping speech because we needed a baseline (i.e., an equivalent nonoverlapping version) to test the results of the source separation models.

2.1 Baseline audio processing

Two speakers were brought into a sound-attenuated booth on different occasions, and each read 300 sentences. One speaker, "Olivia", was a 20-year-old Asian American woman from Atlanta, Georgia who had a relatively high-pitched voice and spoke a variety close to Standard American English. The other, "Tyler", was a 22-year-old White man, also from Atlanta, Georgia, with a relatively low-pitched voice and some features of Southern American English. The sound quality for these two recordings was excellent and had a consistent volume throughout.

To produce a baseline set of formant measurements for these recordings, we processed them using typical procedures. They were manually transcribed at the utterance level in Praat, and were then subjected to force-alignment using the Montreal Forced Aligner (McAuliffe et al. 2017) with vowel formant estimates extracted using FAVE (Rosenfelder et al. 2014), both using the DARLA online web interface (Reddy and Stanford 2015). While there are likely some errors in these measurements because of the automated tools, we used them as a replicable standard to which the test measurements were compared. Prior to processing, we downsampled the audio from 44.1 kHz to 16 kHz because the source separation tools also do such downsampling by default. We note that since the relevant vowel formant data we analyze in this paper falls well below that frequency, we do not expect this to have affected our results in any appreciable way. However, additional work is needed to see how such downsampling affects sounds like [s] that have important high frequency energy.

2.2 Creation of test audio

To create overlapped audio, we took both recordings and merged them into a single mono audio file. However, since both speakers read the same script, they were saying the same sentences around the same time, particularly in the beginning of the file. We did not know how that kind of unison overlap would affect the source separation, so before combining the audio, we swapped the first and second halves of Tyler's recording so that similar sentences would be far from each other in the merged file. We also trimmed the end of Tyler's 36-min recording to match Olivia's 33-min file. The result was rather cacophonous and contained far more overlap than is found in natural conversation or in data normally considered for sociophonetic analysis (around 53.6 % of the audio was overlapping speech). However, we felt it was appropriate, for the purposes of this study, to "stress test" the source separation tools on this kind of data.

2.3 Source separation

We used three open-source models to separate the speech in the artificially overlapped audio file: Libri2mix (https://huggingface.co/speechbrain/sepformer-libri2mix), Whamr16K (https://huggingface.co/speechbrain/sepformer-whamr16k), and WSJ02mix (https://huggingface.co/speechbrain/sepformer-wsj02mix). All three models are based on the transformer-based neural network algorithm SepFormer (Subakan et al. 2020) and were implemented with the SpeechBrain audio toolkit. The models differ only by the training data used to create them (for details, see Cosentino et al. 2020). To facilitate processing the merged audio file, chunks of at most 30 seconds were created from the downsampled audio with the pydub Python module (https://pydub.com). We sought to chunk the audio into more natural breaks that did not interrupt either speaker mid-word, but because of the extreme amount of overlap, simultaneous pauses by both speakers occurred less often than every 30 seconds. For real-world applications of these tools, we recommend breaking at pauses, which would be more frequent in natural recordings. The resulting chunks from each model were then concatenated into a single audio file for each speaker, yielding six new audio files (2 speakers $\times$ 3 SepFormer models).

We then manually spot-checked the six concatenated audio files to identify errors in speaker identification and to subjectively assess the general quality of the source separation process. This was accomplished by listening to small sections of the audio at approximately 30-second intervals and checking for changes in the prominent voice. When encountering a change in speaker, we listened to the full section and noted the beginning and ending timestamps of the aberrant clip. We also made qualitative notes about unexplained extraneous noise, distortion, and the degree to which the other speaker's voice could be heard in the background.

2.4 Additional processing

For each of the six new audio files, we used FAVE to produce formant measurements from an MFA-generated phoneme-level transcription. To ensure identical timestamps across all four versions of the recording, we used the force-aligned TextGrids created from the original audio. Had we force-aligned the separated audio files individually, the aligner likely would have placed boundaries for each vowel in different locations, leading FAVE to extract formants at different time points. Because even small differences in time points can produce unexpectedly different formant measurements (Kendall and Vaughn 2020; Strelluf and Gordon 2024), it would have been impossible to disentangle differences based on audio from differences based on time point.

In the end, we had eight outputs from FAVE: two based on the downsampled original recordings and six based on the source-separated versions of those recordings. For each file, we then followed the order of operations recommended in Stanley (2022). That is, we first classified vowels into major allophonic categories like pre-lateral, pre-nasal, pre-rhotic, and pre-obstruent. Then, for each vowel category, we calculated Mahalanobis distances to detect and remove observations too far from the centroid of their respective cluster (Renwick and Ladd 2016) as a way to remove extreme values that are likely the result of bad acoustic measurements. Normalization was not

necessary because we do not make comparisons across speakers. Finally, we removed formant measurements from pre-sonorant allophones, stopwords, diphthongs, unstressed vowels, and trajectories of each vowel so that we could focus on the midpoints of stressed pre-obstruent monophthongs. This left an average of 1,039 vowel tokens per audio file.

3 Results

3.1 Auditory checks of model performance

The purpose of the auditory checks was to determine potential problems with speaker identification and audio separation. Errors in speaker identification resulted in a change in prominent speaker voice in the concatenated audio file. Problems with audio separation were evidenced by two overlapping voices and/or distortion of the signal. Performance on these tasks varied across models. We outline key differences from the auditory checks here to illustrate the kinds of issues a researcher might encounter when using a source separation tool.

The split files created by Libri2 performed best overall. In these 33-min recordings, there were just six small sections (totaling approximately 2 min 15 sec) for which the tool mislabeled the two speakers, and a couple of places in which a stray syllable or two from Olivia was included with Tyler's speech. However, the separated audio is quite clean, with little noise and very little bleed-through from the other voice. Most of the file with Tyler's speech sounds like a single-speaker recording. The file with Olivia's voice is also good, but the audio sounds more distorted than Tyler's file. The reason for more distortion with Olivia's voice than with Tyler's remains unclear. Whamr produced no obvious errors in speaker identification, though the separated audio files have more audible distortion and noise. WSJ02 produced the poorest results. So much overlapping speech remains in the "separated" audio files that it was sometimes difficult to determine which speaker the tool was targeting. And although the audio is distorted, it is easy to understand what both speakers are saying in both audio files. In other words, the resulting files from WSJ02 do not sound separated at all.

A spectrogram comparison illustrates some of these differences. In Figure 1, we present waveforms and spectrograms for the first sentences of Tyler's speech as isolated by the three models. The box in panel A highlights a brief error in speaker identification, since this bit was said by Olivia. The larger box in panel C identifies a section in which Olivia's voice is more prominent than Tyler's. Arrows in all three panels identify the pauses in speech between sentences and highlight the amount of residual noise and background speech in each audio file.

3.2 Formant comparisons

After extracting formants from each recording, we examined the output by plotting the results in a typical F1-F2 plot (Figures 2 and 3). Note that despite the errors documented in the previous section, we did not do any manual correction after the source separation (i.e., removing stray syllables, fixing "swapped" audio, excluding remaining overlap, etc.). Because the purpose of this case study is to test whether these tools fit into an automated pipeline, we performed no human intervention between steps.

Superficially, we see relatively few differences between the formants produced by the original audio and Libri2- and Whamr-produced audio for both speakers. However, while WSJ02 produced a clean vowel space, the formant measurements' distributions for each vowel are somewhat wider than in the original audio. In the case of Olivia's front and low vowels, there appear to be many formant tracking errors, likely the result of Tyler's voice leaking through. We tentatively conclude, based on these visualizations and the auditory analysis above, that the Libri2mix and Whamr16k models are better than the WSJ02 model in source separation for sociophonetic analysis.

We now move on to a more objective analysis by comparing these distributions more explicitly. For this analysis, we used Strelluf and Gordon's (2024: 54–80) methods as our guide. They compared automatically extracted formant estimates to those with varying degrees of interventions (e.g., manually correcting vowel

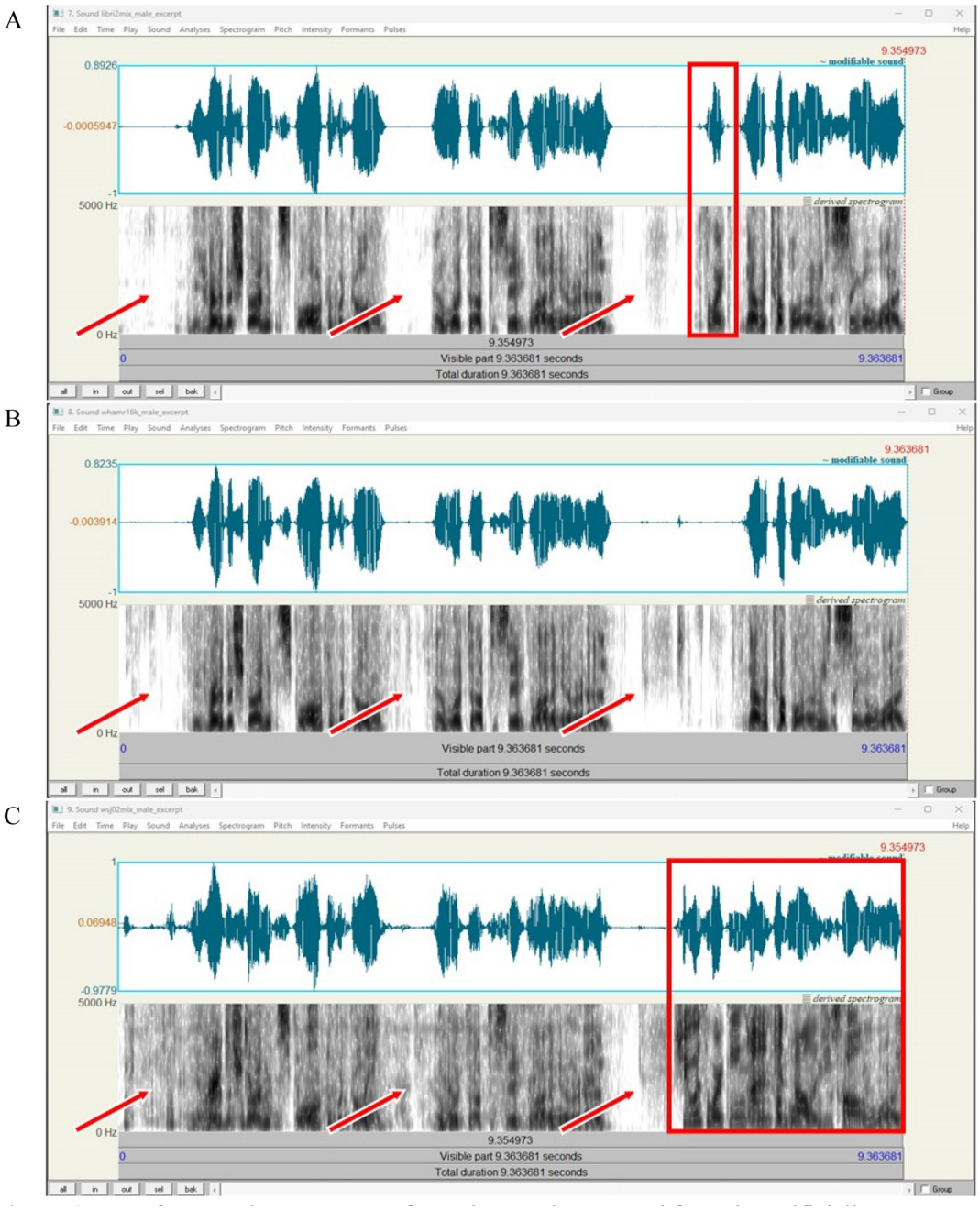

Figure 1: Waveforms and spectrograms for male speech separated from the artificially overlapped audio file using three different models: *A*, Libri2; *B*, Whamr; *C*, WSJ02.

boundaries, manually extracting formant estimates, etc.) using paired Mann–Whitney U tests. This test is appropriate for comparisons of vowel formants because it does not assume normality in the distribution of observations (Levshina 2015: 88). The paired version of this test is especially appropriate since each observation is linked to alternates from the source-separated versions of the audio.

Figure 4 shows the results of these tests. Panels show all 11 monophthongs and are organized by speaker, formant, and model. A perfect source separation model would produce vowel formant measurements that are virtually indistinguishable from the original. In such a case, the *p* value for the statistical test would be high,

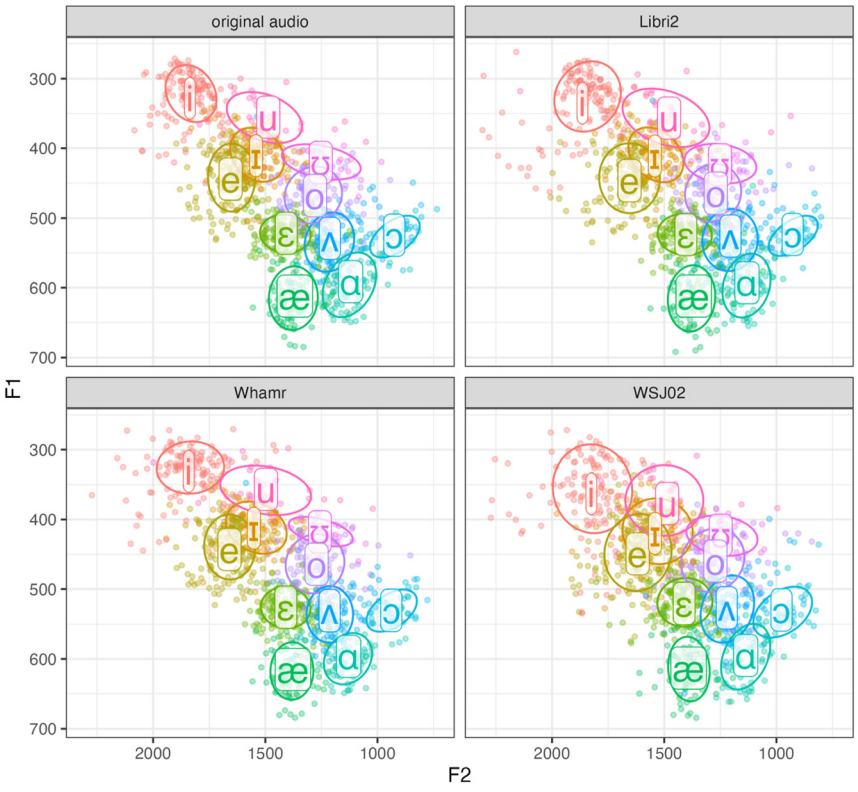

Figure 2: Vowel plots for Tyler.

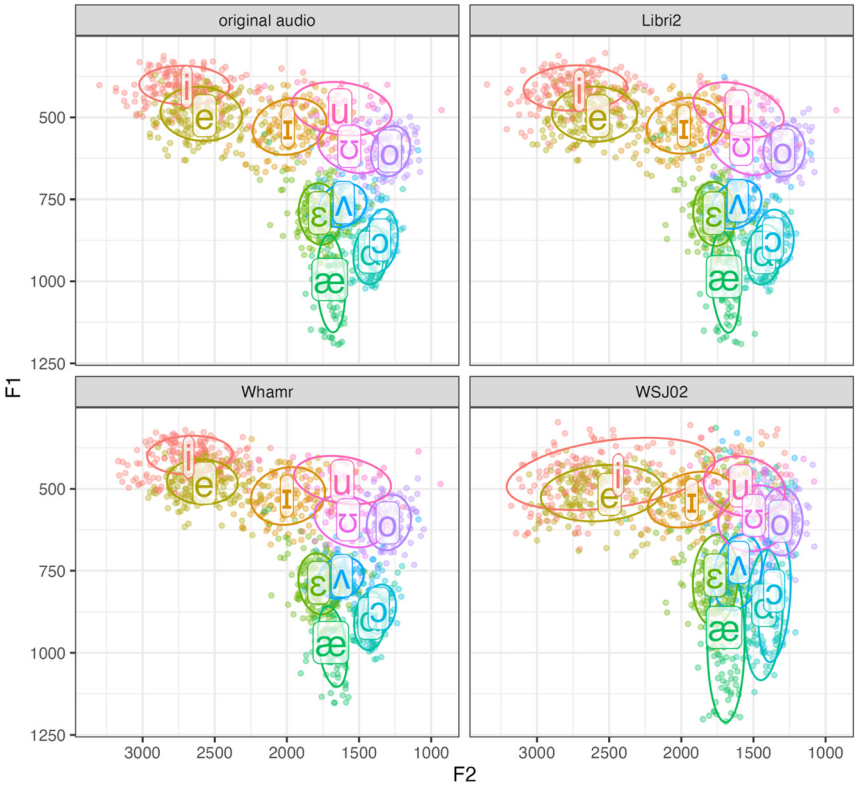

Figure 3: Vowel plots for Olivia.

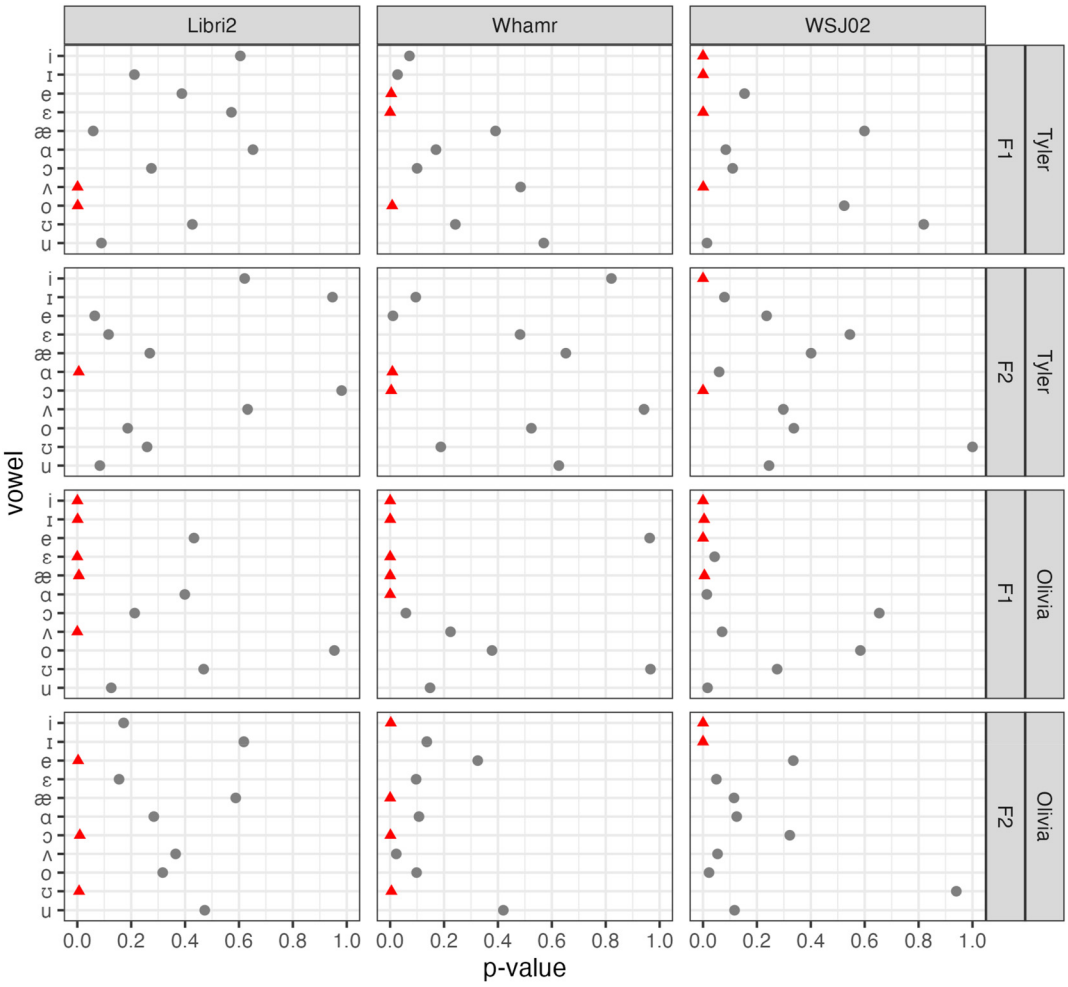

Figure 4: *p* Values from paired Mann–Whitney U tests for each vowel by formant and speaker for each source separation model, when compared to the original. Red triangles indicate statistically significant differences ($p < 0.01$).

ideally close to 1, but certainly greater than 0.05. In our case, because we ran so many statistical tests, we lowered our alpha level to 0.01 (meaning a *p* value should be less than 0.01) to avoid reporting false positives.

We see a concerning number of red triangles in Figure 4. Several statistically significant differences between the original audio and the WSJ02-separated output support the findings reported in Section 3.1 and illustrated in Figures 2 and 3. These differences are primarily found in the F1 of front vowels. All three models produced statistically significant differences for the F1 estimates of Olivia's vowels, primarily her front vowels. Libri2 had the fewest statistically significant differences overall and did particularly well on Tyler's voice.

To illustrate what these comparisons look like at a closer level, Figure 5 shows two example vowels, one with no statistically significant difference and one with a significant difference. Boxplots show the distributions of observations, and lines connect formant measurements from the same time point. The panel on the left shows the F1 of Tyler's /ɑ/, based on the original audio and Libri2. The means and overall distributions are similar. The right panel shows the F1 of Olivia's /i/ from the same model. Olivia's means are still close, even though the test suggests a significant difference. This is encouraging because it suggests that while the difference may be statistically significant, it may not actually make any meaningful difference within the context of formant values.

However, what is more concerning with both of these plots is the number of diagonal lines connecting them. In theory, a perfect model would produce mostly horizontal lines since formant estimates between the two audio files would be nearly the same. While we do see a large number of near-horizontal lines, there are also a fair

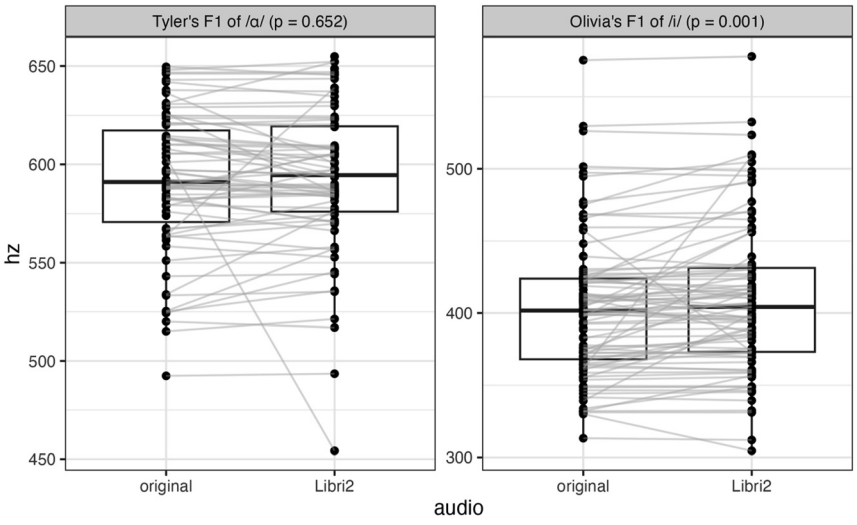

Figure 5: Examples of nonsignificant (*left*) and significant (*right*) pairwise differences (Olivia's data has been downsampled to match the number of observations in Tyler's data to facilitate comparison between the two.).

number of observations connected by steep diagonal lines. This means that some observations were estimated to have drastically different formant measurements in Libri2 compared to the original.

In this case, the F1 of Tyler's /ɑ/ had an average difference of just 1.2 Hz across models. Though, as Strelluf and Gordon (2024: 60) point out, average distances give an idea of how far off the *means* would be in the two analyses. But if we take the absolute value of those differences and average them, the negatives and positives do not cancel out, and we get a more accurate picture of how far off each *individual observation* is likely to be. In this case, the mean absolute value difference is larger at 11.1 Hz. The F1 differences of Olivia's /i/ averaged −6.4 Hz (which is visible in Figure 5), while the mean absolute value of those differences was 12.4 Hz. Kendall and Vaughn (2020) suggest that differences less than 10 Hz for F1 and 15–20 Hz for F2 are within the range of variability for LPC-based formant estimation. While the mean differences in these two examples are less than those thresholds, the mean absolute value of those differences is not.

Table 1 shows the mean and mean absolute value differences from the original for each vowel by model across the entire dataset. Since Strelluf and Gordon's (2024: 61) case study was on a male speaker, they provide no comparable data for Olivia; however, we find that Tyler's differences are generally smaller than Strelluf and Gordon's reported differences. (We note, though, that our data was lab quality and theirs was digitized from a cassette tape.) Comparing our numbers to those reported by Kendall and Vaughn (2020), we see that the actual F1 and F2 differences in our data were lower than their threshold, except for WSJ02 for Olivia for both formants. However, the mean absolute value of differences was lower than their threshold only for the F1s of Tyler's Libri2 and Whamr output.

4 Discussion

In this study, we compare formant estimates from a pair of pristine recordings to merged-and-separated versions of those same recordings. We find that some source separation models are better than others, though even the best models produce statistically significant differences in some cases. While the mean formant measurements per vowel were usually small, differences for each observation were larger in unpredictable ways.

These results are remarkably similar to those found by Strelluf and Gordon (2024: 54–80) in their meta-analysis of sociophonetic data processing. In their comparisons of hand-measured vowel formants to those generated automatically, they find, somewhat paradoxically, that while average formant measurements per vowel were comparable, the measurements for any one vowel token were unreliable. They therefore caution against a word-level analysis or putting too much weight on the formant estimates of any one vowel token since it is influenced in unpredictable ways by the methods taken to get those numbers. In our study, we similarly find

Table 1: Mean differences from the original audio, by formant and model. Columns labeled "actual" show the actual mean of the differences (with pluses and minuses canceling out). Columns labeled "abs" show the mean of the absolute value of differences. The average rows at the bottom of each speaker's table show the mean absolute value of the means for each vowel.

		F1						F2					
		Libri2		Whamr		WSJ02		Libri2		Whamr		WSJ02	
	Vowel	Actual	Abs	Actual	Abs	Actual	Abs	Actual	Abs	Actual	Abs	Actual	Abs
Tyler	i	−0.89	6.64	−0.67	10.00	−35.55	42.69	−2.18	26.62	2.64	51.35	24.59	58.98
	ɪ	−0.14	5.57	−1.65	6.57	−10.15	19.98	0.82	23.64	−10.96	37.96	2.59	65.71
	e	−1.15	5.94	3.41	7.78	1.29	18.07	−2.31	25.92	−11.14	26.35	20.99	57.76
	ɛ	−1.10	4.98	−4.19	8.91	3.08	13.68	2.45	24.41	−1.29	25.07	0.69	37.01
	æ	1.17	8.91	−2.38	11.91	−0.49	14.15	−2.93	29.81	−11.96	32.41	−22.00	47.02
	ɑ	1.28	11.12	−5.02	13.95	3.04	16.88	−8.92	18.87	−9.87	26.09	−10.69	32.80
	ɔ	1.08	9.03	−4.62	11.53	−4.60	16.56	−2.67	16.79	−8.79	21.17	−47.03	57.79
	ʌ	3.24	7.82	−2.20	10.91	5.00	23.70	7.67	23.83	−1.97	20.43	−8.67	38.87
	o	3.40	5.17	4.97	8.21	0.32	14.64	−3.99	16.07	0.76	11.69	9.44	37.97
	ʊ	−2.82	9.05	2.52	8.27	−2.18	15.15	−9.68	33.02	−14.60	34.84	4.13	43.49
	u	2.68	6.10	−1.30	6.47	−18.06	28.44	14.55	56.91	−10.12	49.79	6.46	51.36
	Mean (abs)	1.72	7.30	2.99	9.50	7.61	20.36	5.29	26.90	7.65	30.65	14.30	48.07
Olivia	i	−6.43	12.44	8.49	16.79	−44.11	55.77	4.75	73.98	24.67	89.35	238.50	309.57
	ɪ	2.02	7.38	7.60	14.18	−5.33	22.52	7.53	21.64	−7.76	34.70	63.65	95.49
	e	−1.29	5.65	1.89	10.06	−10.92	20.43	3.83	45.78	1.04	45.07	56.25	102.46
	ɛ	2.16	10.79	8.14	17.39	30.98	47.67	−1.04	20.94	−3.29	22.98	6.35	30.71
	æ	7.27	17.43	27.07	34.36	84.99	97.22	−2.47	12.13	9.39	25.94	13.27	37.96
	ɑ	3.78	13.42	18.84	30.31	51.25	69.27	−1.82	10.34	−2.82	15.11	−4.83	21.54
	ɔ	−6.61	16.06	12.38	25.64	55.93	94.82	−6.34	14.68	−11.10	15.98	−4.46	29.74
	ʌ	10.50	14.26	3.29	15.92	35.19	49.65	−1.26	21.25	−6.16	16.67	6.40	32.00
	o	−1.29	9.70	2.61	13.32	9.37	23.47	−0.99	6.85	−5.06	18.07	−8.47	12.69
	ʊ	5.44	14.63	−1.91	12.21	15.59	28.47	−9.94	15.61	−12.34	23.12	61.26	81.54
	u	−3.58	11.61	−1.35	7.88	−16.88	40.26	15.66	50.71	15.84	70.26	43.78	132.26
	Mean (abs)	4.58	12.12	8.51	18.01	32.78	49.96	5.06	26.72	9.04	34.30	46.11	80.54

that the differences in mean vowel formant measurements were small, particularly with the Libri2mix and Whamr16k models. Even though some differences were statistically significant, the magnitude of those differences was usually within the range of formant estimation variability anyway (Kendall and Vaughn 2020), meaning we can cautiously disregard the output of those statistical tests. At the individual token level, the differences between our recordings appear to be smaller than Strelluf and Gordon's (2024) reported differences, suggesting that the effect of source separation on formant measurements may be smaller than the interventions they experimented with (hand-correcting automatically generated vowel boundaries, correcting outliers, post hoc recoding of lexical sets).

These results are encouraging. Even on extreme amounts of overlap, the Libri2 model did very well, with Whamr close behind it, at source separation and recovering the original audio. In cases where the original audio has overlap, such as in a sociolinguistic or other kinds of interviews, these models may reliably and accurately separate speakers into their own tracks, allowing researchers to analyze otherwise irrecoverable linguistic data. Future research would do well to test this hypothesis.

4.1 Application to real-world overlap

Before we recommend that sociophoneticians (especially historical sociophoneticians) use these source separation models on their data, it is worth acknowledging several primary differences between our case study and real-world overlap.

First, our recordings were artificially overlapped, and while the difference between artificial and natural overlap itself may not be meaningful, the amount of overlap certainly is. Real-world data almost certainly has less overlap than was present in our test file. This means that the effect that source separation has on the overall results will be much smaller than what is reported here, since it affects a smaller proportion of the data.

However, there are several caveats. Our study only examined a case where there were exactly two speakers. Source separation models can work with a greater number of speakers, even if the precise number is not known, but we assume that in such cases the output will be less clean. The audio we used was also pristine. It was recorded in a lab with high-quality equipment under ideal circumstances. It is unlikely that sociophonetic data, such as that found in modern-day interviews or that recovered from archival sources, will be so clean. We assume audio quality and background noise would lower the overall effectiveness of source separation, but future work should analyze these factors. The two original files in our study were about equal volume as well. Interviews conducted with just one microphone placed closer to one speaker will result in one speaker coming in more quietly than the other. We are not sure whether unequal speaker amplitudes will decrease the quality of the separation or, perhaps because different amplitudes will make the speakers more different from each other, increase it. That said, as a preprocessing step, a researcher could apply an audio compressor to even out different levels of volume in the recording before applying source separation.

Our speaker dyad included two very different voices. One speaker had a relatively higher-pitched voice and a shorter vocal tract while the other was lower-pitched with a longer vocal tract. This may have made it easier for the source separation models to do their job. We pointed out already that even with pauses interspersed, rarely did the Libri2 and Whamr models switch from one speaker to another. We suspect that two speakers with more similar-sounding voices will not be separated quite as cleanly as our two speakers were. We encourage additional research to test how voice similarity affects source separation, and we encourage those who need to implement these methods on more similar-sounding dyads to be more cautious when evaluating the output.

On a related note, as mentioned above, WSJ02 did the poorest job at source separation and quite a lot of one speaker's audio bled into the other's. The wide distribution of the F2 of Olivia's front vowels likely is the result of this bleeding, since some measurements were influenced by relics of Tyler's lower voice. Same-gender dyads may result in superficially cleaner output. However, if there is as much bleeding in those dyads as there was in our study, the result may be a mix of vowel formant measurements from both speakers. And since the two voices are similar, it will be much harder to detect errors. We therefore recommend checking how much separation actually occurs before proceeding with sociophonetic analysis.

4.2 Recommendations

We are cautiously optimistic about source separation and its application in sociophonetics. The results of this limited and controlled test are encouraging. In cases where there are only two speakers with sufficiently different voices, the audio quality is good, and there is limited overlap, source separation will likely do a good job. We end with the following recommendations for those interested in using source separation models in sociophonetic data processing:

- Experiment with different models to find the one that is most suitable. In this rapidly developing field, models improve over time and new options emerge. So, rather than clinging to one that works, continuously explore potentially better tools.
- Split audio at natural breaks rather than equal intervals. This can be done manually or using a script.
- Listen to the output to ensure clean separation.
- Transcribe based on the source-separated audio or ensure that existing transcriptions still match the new audio before conducting acoustic analysis.
- Treat formant estimates at the token level with caution. To be safe, only do analyses on vowel summaries like averages.
- Carefully document and report all methodological choices and human interventions.

We encourage additional research on source separation for linguistic studies and hope to see it used for recovering previously unanalyzable audio.

References

Barreda, Santiago. 2021. Fast Track: Fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard* 7(1). 20200051.

Boudahmane, Karim, Mathieu Manta, Fabien Antoine, Sylvian Galliano & Claude Barras. 1998. Transcriber. Available at: http://trans.sourceforge.net/.

Bowie, David. 2003. Early development of the card-cord merger in Utah. *American Speech* 78(1). 31–51.

Brugman, Hennie & Albert Russel. 2004. Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation. Lisbon, 26–28 May*. http://lrec-conf.org/proceedings/lrec2004/ (accessed 6 January 2025).

Cheng, Andrew. 2018. A longitudinal acoustic study of two transgender women on YouTube. *UC Berkeley Phonology Lab Annual Reports* 14. 168–188.

Cheng, Andrew. 2023. Second dialect acquisition "in real time": Two longitudinal case studies from YouTube. *American Speech* 98(2). 194–224.

Cosentino, Joris, Manuel Pariente, Samuele Cornell, Antoine Deleforge & Emmanuel Vincent. 2020. LibriMix: An open-source dataset for generalizable speech separation. *arXiv*. http://arxiv.org/abs/2005.11262.

Gorman, Kyle, Jonathan Howell & Michael Wagner. 2011. Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39(3). 192–193.

Harrington, Jonathan, Sallyanne Palethorpe & Watson Catherine. 2000. Monophthongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association* 30(1–2). 63–78.

Hickey, Raymond (ed.). 2017. *Listening to the past: Audio records of accents of English* (Studies in English Language). Cambridge: Cambridge University Press.

Holliday, Nicole. 2024. Complex variation in the construction of a sociolinguistic persona: The case of Vice President Kamala Harris. *American Speech* 99(2). 135–166.

Kendall, Tyler & Charlotte Vaughn. 2020. Exploring vowel formant estimation through simulation-based techniques. *Linguistics Vanguard* 6(s1). 20180060.

Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347.

Lee, Sarah. 2017. Style-shifting in vlogging: An acoustic analysis of "YouTube Voice". *Lifespans and Styles* 3(1). 28–39.

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Ma, Marcus, Lelia Glass & James Stanford. 2024. Introducing Bed Word: A new automated speech recognition tool for sociolinguistic interview transcription. *Linguistics Vanguard* 10(1). 641–653.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association [Interspeech]*, 498–502. Stockholm, Sweden.

Mendoza-Denton, Norma. 2011. The semiotic hitchhiker's guide to creaky voice: Circulation and gendered hardcore in a Chicana/o gang persona. *Journal of Linguistic Anthropology* 21(2). 261–280.

Mikolov, Tomas, Kai Chen, Greg Corrado & Dean Jeffrey. 2013. Efficient estimation of word representations in vector space. *arXiv*. http://arxiv.org/abs/1301.3781.

Olsen, Rachel M., Michael L. Olsen, Joseph A. Stanley, Margaret E. L. Renwick & William A. Kretzschmar Jr. 2017. Methods for transcription and forced alignment of a legacy speech corpus. *Proceedings of Meetings on Acoustics* 30(1). 060001.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey & Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*. Honolulu, HI.

Reddy, Sravana & James N. Stanford. 2015. Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard* 1(1). 15–28.

Renwick, Margaret E. L. & D. Robert Ladd. 2016. Phonetic distinctiveness vs. lexical contrastiveness in non-robust phonemic contrasts. *Laboratory Phonology* 7(1). 1–29.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard & Jiahong Yuan. 2014. FAVE (Forced alignment and vowel extraction) program suite, version 1.2.2. Available at: https://doi.org/10.5281/zenodo.22281.

Schiel, Florian. 1999. Automatic phonetic transcription of non-prompted speech. In *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco: University of California. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0607.pdf (accessed 7 January 2025).

Stanley, Joseph A. 2022. Order of operations in sociophonetic analysis. *University of Pennsylvania Working Papers in Linguistics* 28(1). Available at: https://repository.upenn.edu/pwpl/vol28/iss2/17.

Strelluf, Christopher & Matthew J. Gordon. 2024. *The origins of Missouri English: A historical sociophonetic analysis*. Lanham: Lexington Books.

Subakan, Cem, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi & Jianyuan Zhong. 2020. Attention is all you need in speech separation. *arXiv*. https://doi.org/10.48550/arXiv.2010.13154.

Wolfram, Walt, Caroline Myrick, Jon Forrest & Michael J. Fox. 2016. The significance of linguistic variation in the speeches of Rev. Dr. Martin Luther King Jr. *American Speech* 91(3). 269–300.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/lingvan-2024-0152).