Comparing Joint Speech and Synchronous Speech: What Happens When We Add More Speakers?

Fred Cummins

School of Computer Science
University College Dublin
fred.cummins@ucd.ie

Abstract. Eight subjects were recorded speaking texts in synchrony in groups of 1, 2, 4, 6 and 8 speakers at a time. The question asked was whether either speech timing or inter-speaker synchronisation would be manifestly different as more speakers speak together. Results showed no appreciable changes to overall speech timing as number of speakers varied. Examination of asynchrony scores likewise showed no effect of number of speakers, even as texts from very different genres were used. The results suggest that synchronous speaking in a laboratory is importantly different from joint speech found in ritual, protest, and elsewhere.

Keywords: Joint Speech · Synchronisation · Synchronous Speech

1 Background

Joint speech is found wherever multiple people utter the same words at the same time [7, 3]. This empirically grounded definition serves to pick out many important domains of human activity, including ritual and prayer, protest, the activities of many sports fans, and a variety of practices in early education. It is thus a fundamental part of human social activity and is found in every known culture and throughout history. Its thematisation as an object of scientific inquiry is, however relatively recent [6]. When faced with an underexplored topic that extends into many forms of culturally saturated collective activities, it is something of a creative challenge to find experimental and modelling avenues that can illuminate aspects of the activity that make human behaviour and experience within these domains intelligible.

One experimental avenue that has produced some limited insights is provided by the vehicle of synchronous speech [1]. This is a joint speaking task conducted in laboratories, in which volunteers are asked to read texts selected by the experimenter with the express goal of speaking in synchrony with one another. Many of the more interesting characteristics of joint speech do not survive translation to the constrained setting of the laboratory. The passion that inspires the protesters, the piety of the faithful, or the enthusiasm of the soccer supporter cannot be reproduced at will by reading unmotivated texts. However,

the ability to remain in time with other speakers is interesting in its own right, and this, more mechanical, aspect of joint speech lends itself to study using the synchronous speech approach.

Past work on synchronous speaking has used dyads (pairs of speakers) exclusively. The body of findings includes the observation that speakers can synchronise effectively and without practice [3]. This is interesting in its own right as speaking is an intrinsically plastic activity that adapts to context and to co-speakers effortlessly, yet in the laboratory, such expressive and context-sensitive variability is discarded without any obvious effort. In the constrained situation of a synchronous speech experiment, an average asynchrony of about 40 ms. is typical. Asynchrony after a long pause is slightly greater, but collapses back to the 40 ms. level within a syllable or two. If speakers are back to back and thus cannot see each other, asynchrony at phrase starts is slightly elevated by about 10 ms. but again, it quickly relaxes to the same values. The synchrony elicited by this experimental methodology is not strictly comparable to that found in the wild, e.g. during group prayers or in the street. One way in which it differs is in the production of a specific kind of speech error unique to the synchronous speaking situation. It has been repeatedly observed that a small hesitation or error on the part of one speaker can frequently lead to an abrupt and simultaneous cessation of speaking by both speakers, which is a kind of production error only found under these conditions. The presence of this error suggests that the two speakers are tightly coupled, and thus non-independent, when performing the task [6]. This is evidence for the existence of an interpersonal synergy [12, 10]. Coupling between speakers has also been documented in an fMRI study in which cortical activity during live synchronisation with an experimenter was interestingly differentiated from activity when speaking along with a recording of the experimenter, even when the subjects were unaware that there was a difference in conditions, or that recordings were being used at all [9].

When joint speech is encountered, there are typically very many more than two speakers involved. Synchronisation is much more lax than found in these dyadic synchronous speech experiments, and, of course, the motivation and context are radically different. In what follows, a small experiment is reported that employs the synchronous speech methodology, but extends it to more speakers, allowing a comparison of the temporal alignment of 2, 4, 6 and even 8 speakers at a time. We pose two substantial research questions. Will the addition of more speakers lead to systematic changes in the overall temporal patterning of speech? And will the addition of more speakers lead to less overall synchrony within the group? This will allow us to better understand the difference in group performance between the experimental situation of synchronous speaking and the collective behaviour of joint speaking in the wild. We know that prosody in chanting is often substantially altered, but this may be a function of repetition, which is frequently present, or of associated body movements, such as fist pumping, or of many other potential causes. We do not know if it is a necessary effect of group synchronisation, though observation of joint speech done instrumentally, such as in public swearing of an oath of allegiance, typically exhibits

less prosodic stylisation than well practiced pledges, chants, and prayers. We also know that synchrony among speakers in ritual and protest is typically not as tight as that found under laboratory conditions of synchronous speaking.

2 Methods

Eight speakers took part (age range: 21–46, 2 Female, 6 Male). All were student volunteers and native speakers of English. No reward was given for participation and all experimental procedures were approved by the UCD research ethics committee. Recording was done in a single class room (approx 3 m x 4 m) and subjects each wore head mounted microphones to minimise cross-speaker contamination. Each subject was recorded on a spearate channel. An ambient recording of all speakers was also captured with a separate microphone.

Five texts were prepared that spanned genres. They included one regular and one irregular wordlist, the Hail Mary prayer, a short prose piece with irregular stressed syllable onsets, and a metrically complex poem. All texts had been used in previous synchronous speech experiments. There were 40 trials. On each trial, 1, 2, 4, or 6 subjects were randomly selected, or all 8 subjects took part, and the 5 texts were read in randomized order. This provided eight complete recordings of all texts for each fixed number of speakers. Speaking began on a signal from the experimenter. No tempo instructions or other guidance was given.

There are several ways to estimate synchrony across any given pair of speakers [4]. Perhaps the simplest is to identify the point at which the intensity rise in a strong stressed syllable occurs (this is an estimate of a P-centre [13]), and to use those as landmarks whose time of occurrance can be compared. Stressed syllable onsets were computed for all recordings and texts using the algorithm reported in [8].

3 Results

Initial results will be presented here for the first time. A first question to be asked is whether overall speech timing is affected by the constraint of speaking in groups of different size. To pursue this issue, we first examine the intervals between stressed syllable onsets in the two wordlists. The regular word list consisted of eight trochees (*Borrow, Dancer, Butter, Dagger, Boiler, Doggie, Body, Deeper*) while the irregular wordlist had a similar strong-weak alternation, but with an irregular patterning of word edges (*Debug, Debasing, Ball, Degree, Bandana, Beach, Beginning, Duck*).

Fig. 1 shows interval durations for the two wordlists as a function of position in list (x-axis) and speaker numbers. Each list displays idiosyncratic timing properties that are highly preserved across changes to speaker number, and, crucially, there is no evidence here for any systematic change as speaker numbers vary from 1 to 8. The alternating pattern of relatively shorter and longer intervals is of a kind with previously presented results from both speech and typing [2, 11].

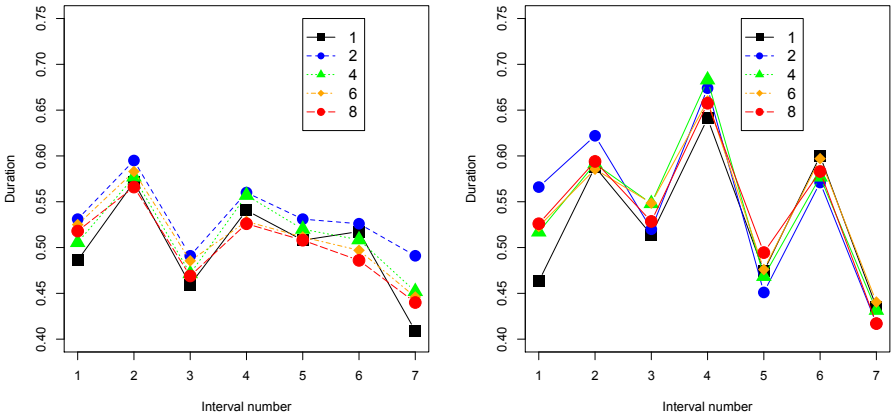

Fig. 1. Inter-onset interval duration (seconds) as a function of interval number (x-axis) and speaker number (legend). Left: intervals from onsets of 8 successive trochees. Right: intervals from a wordlist with similar alternation, but with irregular word boundaries.

We repeat the analysis now with a much more rhythmically complex text. The poem "Kill a cat" displays a binary rhythm in lines 1, 2, and 5, but switches to a triple metre in lines 3 and 4[1]. The text is:

Kill a cat, kill a cat
Bash its brains in with a bat.
Their nine lives expire
When tossed in a fire, so
Kill a cat today.

Fig. 2 shows boxplots of interval duration between onsets of stressed syllables for the two conditions of $n = 1$ and $n = 8$. All intermediate cases ($n = 2, 4, 6$) were also examined and found not to differ significantly. The metrical pattern of the poem is clearly evident in these two limiting cases, with the switch to triple metre generating longer intervals between stressed syllable onsets. But once, more, there is no evidence for any appreciable reorganisation or prosodic modification as the number of speakers rises. (Note: inferential statistical tests would be highly inappropriate here, given the complexity of the data.)

We now move to examine asynchrony, computed by pairwise computation of lags between corresponding points in two waveforms. Overall the distribution of lags is strongly right skewed, so all lags were log transformed, resulting in an approximately normal distribution of lags overall. Fig. 3 shows lags as a function of text and number of speakers for all five texts. Note that because lags

[1] Apologies to cat lovers everywhere.

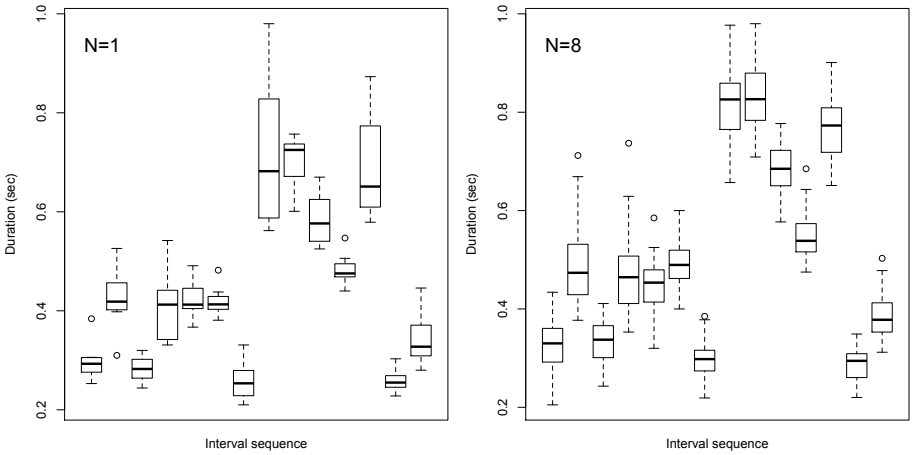

Fig. 2. Inter-stressed onset interval duration (seconds) in a metrically complex poem. Left: single speakers. Right: Groups of 8 speakers.

are computed on a pairwise basis, there are many more pairwise comparisons when $n = 8$ than when $n = 2$ (224 vs. 8).

From the figure it is clear that asynchrony remains more or less constant as the number of speakers increases, irrespective of text type. There is some evidence that the simpler texts (word lists) result in slightly tighter synchrony than the complex poem or prose passages, but the principal result here is a lack of effect of the number of speakers.

4 Discussion

Joint speech and synchronous speech are not the same thing. The first is found in a wide variety of situations in which collective purposes, passions, and aspirations are made manifest, and in which different kinds of identies are brought forth or enacted, whether that be as Buddhist monks, concerned citizens, or Arsenal supporters. The latter is found in the laboratory, and although there too people speak in synchrony with one another, their purposes are starkly different, and the speech produced has its own specific characteristics. It usually exhibits a very strong synchrony that is alien to joint speech in the wild. But the circumstances under which joint speech is produced vary greatly and they frequently include contextual factors and historical particulars that leave their stamp on the speech. Thus prayers and ritual incantations are typically repeated many times over. Protest chants are repeated and fequently augmented with fist pumping, clapping or noise making of various kinds. Football chants too typically involve repetition and musical elements. (In studying joint speech, a determinate border between speech and song or music can no longer be sustained [5].)

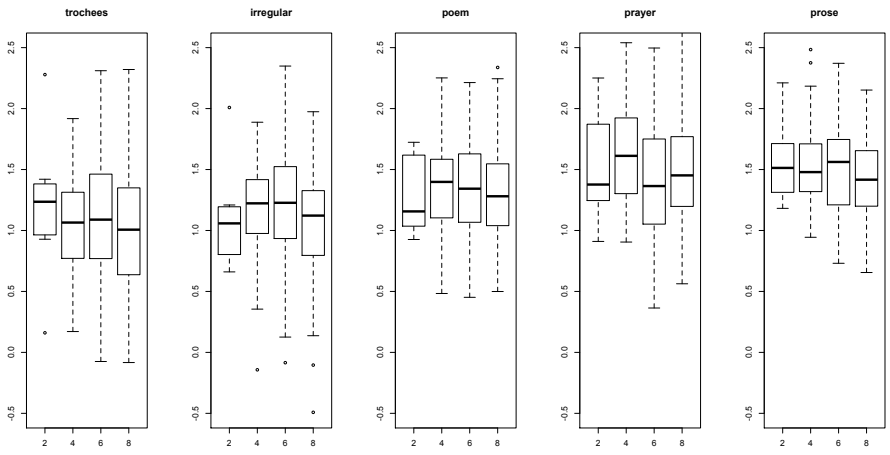

Fig. 3. Log transformed lags between corresponding points in parallel waveforms as a function of text and number of speakers.

A significant difference between the context in which joint speech and synchronous speech occur is the restriction of all previous synchronous speech experiments to dyads, whereas joint speech typically involves many more speakers. The present work thus served to see whether this factor, number of speakers, can account for any of the observed differences between the two kinds of speech. The result is in the negative, both when overall temporal patterning is examined, and when pairwise asynchronies are computed. Although a negative result, this is a finding of some significance, as it allows us now to more clearly distinguish between the two kinds of speaking, and to recognise that synchronous speaking is not free of its own contextual imprematur that cannot be attributed simply to the smaller number of participants.

One important lesson to be learned from this small experiment is that the study of joint speech demands a great deal of sensitivity to the context in which it occurs. In keeping with a general trend in phonetic and linguistic research, it is necessary to recognise that the laboratory is not a neutral space in which a pristine object of study, speech, can be found. It is, rather, itself a rich context imbued with its own characteristics that indelibly mark the speech produced therein. The enhanced synchrony found in synchronous speech, and the associated characteristic phenomenon of joint cessation when an error occurs, these are not a result of a lack of numbers, but a product of an altered situation of speaking.

References

1. Cummins, F.: On synchronous speech. Acoustic Research Letters Online **3**(1), 7–11 (2002). https://doi.org/doi = 10.1121/1.1416672

2. Cummins, F.: Rhythmic organization of read word lists. Journal of the Acoustical Society of America **112**(5:2), 2443 (2002)
3. Cummins, F.: Practice and performance in speech produced synchronously. Journal of Phonetics **31**(2), 139–148 (2003)
4. Cummins, F.: Measuring synchronization among speakers reading together. In: Proc. ISCA Workshop on Experimental Linguistics. pp. 105–108. Athens, GR (2006)
5. Cummins, F.: Joint speech: The missing link between speech and music? Percepta—Revista de Cognição Musical **1**(1), 17–32 (2013)
6. Cummins, F.: The remarkable unremarkableness of joint speech. In: Proceedings of the 10th International Seminar on Speech Production. pp. 73–77. Cologne, DE (2014)
7. Cummins, F.: The Ground From Which We Speak: Joint Speech and the Collective Subject. Cambridge Scholars (2018)
8. Cummins, F., Port, R.F.: Rhythmic constraints on stress timing in English. Journal of Phonetics **26**(2), 145–171 (1998)
9. Jasmin, K.M., McGettigan, C., Agnew, Z.K., Lavan, N., Josephs, O., Cummins, F., Scott, S.K.: Cohesion and joint speech: Right hemisphere contributions to synchronized vocal production. The Journal of Neuroscience **36**(17), 4669–4680 (2016)
10. Riley, M.A., Richardson, M., Shockley, K., Ramenzoni, V.C.: Interpersonal synergies. Frontiers in Psychology **2**, 38 (2011)
11. Rosenbaum, D.A., Kenny, S.B., Derr, M.A.: Hierarchical control of rapid movement sequences. Journal of Experimental Psychology: Human Perception and Performance **9**(1), 86–102 (1983)
12. Schmidt, R.C., Carello, C., Turvey, M.T.: Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. Journal of Experimental Psychology: Human Perception and Performance **16**(2), 227 (1990)
13. Scott, S.K.: P-centers in Speech: An Acoustic Analysis. Ph.D. thesis, University College London (1993)