These datasets comprise listeners' transcriptions of sentence-length speech analogues, and the scores and error counts derived from them, for Experiments 1 and 2 of the article of the same title (Summers and Roberts, 2020; Journal of the Acoustical Society of America).

Experiment 1:

There is one spreadsheet for experiment 1, comprising a summary worksheet and the raw data for each listener. The summary worksheet contains aggregated scores (proportion keywords correct by tight scoring, see below) for each listener in each condition, with relevant demographic information. For completeness, the summary worksheet also contains the mean number of words reported by each listener per condition. Subsequent worksheets comprise the raw data for each listener and stimulus.

The raw data comprise: (a) the stimulus presented [Column heading: Stimulus], (b) the stimulus sentence [Text], (c) the listener's response [Transcription], (d) the condition number for which the stimulus was presented [Condition], (e) the number of times the listener heard the stimulus (always once in this experiment), (f) the number of keywords in the stimulus, (g) the loose score (number of keywords reported correctly for which the stem of the word is correct – e.g., "type", "types", and "typed" would all be marked correct for keyword "typing"; the loose score was not analysed but is included for completeness), and (h) the tight score (only exactly reported keywords are marked as correct; homonyms are accepted). The mean scores for each condition – (i) loose and (j) tight – are computed by dividing the number of correct keywords reported for all sentences in the condition (6 sentences/condition) by the total number of keywords. Additionally, column (l) [Number of words reported] computes the number of words transcribed by the listener regardless of their correctness and (m) computes the mean number transcribed for each condition.

Experiment 2:

There are two spreadsheets for experiment 2, one for the main experiment and one for the follow-up experiment.

The spreadsheet for the main experiment comprises three summary worksheets and the raw data for each listener. The SummaryTarget worksheet contains aggregated scores (proportion keywords correct by tight scoring) for each listener in each condition (C1-C7) and also collapsed across target-to-masker ratio for the conditions involving interferers (C1-C3 vs C4-C6), with relevant demographic information. The SummaryInterferer worksheet contains aggregated scores (proportion of keywords reported that are exclusively in the interferer by tight scoring). Finally, the SummaryError worksheet contains the count of two types of error – intrusion errors and other errors (for both loose and tight scoring). Intrusion errors are the number of keywords reported that were exclusively in the interferer(*). Other errors are the number of words reported that would usually be classified as keywords but which were not present in either the interferer or the target.

(*) Owing to small differences in the number of keywords per asynchrony condition (54 or 55), note that the correspondence between the proportion of keywords reported that are exclusively in the interferer (SummaryInterferer) and the intrusion error count (SummaryError) is close but not exact.

Subsequent worksheets comprise the raw data for each listener and stimulus. The raw data comprise: (a) the stimulus presented [Column heading: Stimulus], (b) the listener's response [Transcription], (c) The text of the target sentence [TargetText], (d) the text of the interferer sentence [InterfererText] where applicable, (e) the condition number for which the stimulus was presented [Condition], (f) the number of times the listener heard the stimulus (always once in this

experiment), (g) the number of keywords in the target, (h) the loose score (number of target keywords reported correctly for which the stem of the word is correct – e.g., "type", "types", and "typed" would all be marked correct for keyword "typing"), (i) the tight score (only exactly reported target keywords are marked as correct; homonyms are accepted), (j) the number of keywords in the interferer, (k) the number of words reported that were exclusively in the interferer according to loose scoring, (l) the number of words reported that were exclusively in the interferer according to tight scoring, (m) the number of intrusion errors based on loose scoring [ErrorIntrusionLoose], (n) the number of words reported that were not in either the interferer or the target based on loose scoring but which would otherwise be classified as keywords [ErrorOtherLoose], (o) the number of intrusion errors based on tight scoring [ErrorIntrusionTight], (p) the number of words reported that were not in either the interferer or the target based on tight scoring but which would otherwise be classified as keywords [ErrorOtherTight], (r) the mean tight scores for each condition are computed by dividing the number of correct keywords reported for all sentences in the condition (6 sentences/condition) by the total number of keywords, (t) the mean interferer scores for each condition (tight scoring) are computed by dividing the number of keywords reported that appeared exclusively in the interferer for all interfering sentences in the condition (6 sentences/condition) by the total number of interferer keywords. The final four columns – (v) [SumErrorIntrusionLoose], (w) [SumErrorOtherLoose], (y) [SumErrorIntrusionTight], and (z) [SumErrorOtherTight] – are the total number of those errors for each condition.

The spreadsheet for the follow-up experiment comprises a summary worksheet and the raw data for each listener. The summary worksheet contains aggregated scores (proportion keywords correct by tight scoring, see below) for each listener in the two formant-asynchrony conditions – C1-C3 (0 ms) vs C4-C6 (150 ms) – and the mean number of words reported by each listener per asynchrony condition. Subsequent worksheets comprise the raw data for each listener and stimulus.

The raw data comprise: (a) the stimulus presented [Column heading: Stimulus], (b) the stimulus sentence [Text], (c) the listener's response [Transcription], (d) the condition number for which the stimulus was presented [Condition] (1 to 3 correspond to 0-ms formant asynchrony and 4-6 correspond to 150-ms formant asynchrony), (e) the number of times the listener heard the stimulus (always once in this experiment), (f) the number of keywords in the stimulus, (g) the loose score (number of keywords reported correctly for which the stem of the word is correct – e.g., "type", "types", and "typed" would all be marked correct for keyword "typing"; the loose score was not analysed but is included for completeness), and (h) the tight score (only exactly reported keywords are marked as correct; homonyms are accepted). The mean scores for each asynchrony condition – (i) loose and (j) tight – are computed by dividing the number of correct keywords reported for all sentences in the condition (6 sentences/condition, corresponding to 18 sentences for each formant asynchrony) by the total number of keywords. Additionally, column (l) [Number of words reported] computes the number of words transcribed by the listener regardless of their correctness, and (m) computes the mean number transcribed for each asynchrony condition (C1-C3 vs C4-C6).