



# Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with respect to Turn-taking Activity

Nick Campbell<sup>1</sup>, Stefan Scherer<sup>2</sup>

<sup>1</sup> Centre for Language and Communication Studies, Trinity College Dublin, Ireland

<sup>2</sup> Institute of Neural Information Processing, University of Ulm, Germany

nick@tcd.ie, stefan.scherer@gmail.com

## Abstract

This paper describes a system for predicting discourse-role features based on voice-activity detection. It takes as input a vector of values extracted from conversational speech and predicts turn-taking activity and active-listening patterns using an echo-state network. We observed evidence of frame-attunement using a measure of speech density which takes the ratio of speech to non-speech behaviour per utterance. We noted a synchrony of utterance timing and modelled this using the ESN. The system was trained on a subset of data from 100 telephone conversations from the 1,500-hour JST Expressive Speech Processing corpus, and predicts the interlocutor's timing behaviour with an error-rate of less than 15% based on one partner's speech-activity alone. An integrated system with access to content information would of course perform at higher rates.

**Index Terms:** speech activity, active-listening, synchrony, speech-timing, speech density, feedback.

## 1. Introduction

“Among the key concerns for any study of the relationship between language use and social life is the question of how humans establish what Goffman referred to as ‘mutually ratified participation’ in particular, concrete, situated interactions” [1]

Drawing in part on the later work of Goffman [2], Kendon has describes a process of ‘frame attunement’ as “the mechanism for the mutual coordination of locally relevant situated identities, together with the conjoint establishment of a local space or arena within which a specific interactional encounter is enabled to take place” [3].

Hutchby claims this mechanism to be an incessant accomplishment in human interaction: “It is not enough simply to establish a mutually oriented-to spatial or environmental frame for the encounter at the outset and then proceed. Rather, interactants constantly attune and re-attune their frames according to contingencies of the moment” [1]. In this paper we examine a corpus of telephone speech and show evidence for this process of constant re-attunement.

The paper extends previous work [4] showing that simple measures of speech timing or ‘discourse flow’ can be used to indicate the types of social relationships established between partners in a dialogue. The paper supports Kendon’s view that this is more a matter of spatial orientation than propositional content of the talk, and by extending spacial orientation into the fourth dimension shows how the precise *timing* of utterance segments in a discourse is finely tuned and coordinated in a synchronised manner by the dialogue participants.

## 1.1. Synchrony & Alignment

Previous work [5] on active listening and synchrony in spoken dialogues was based on both telephone speech and material derived from multimodal in-situ recordings of round-table multi-party conversational interactions, showing that participants engage positively in a discourse by synchronising their speech and movements to a very high degree, and frequently speaking and moving simultaneously at points of high engagement. Findings based on analysis of two-party telephone conversations (in Japanese) were confirmed in the multi-party round-table conversation data (in English). The present paper returns to the telephone conversations for a finer and more detailed analysis of timing variation and entrainment *within* turns.

## 1.2. Materials for the Analysis

The data underlying this study are taken from a corpus of recorded telephone conversations [6]. One hundred 30-minute telephone conversations were recorded over a period of several months, with paid volunteers coming to a building in Western Japan once a week to talk with specific partners in a separate part of the same building over an office telephone. While talking, they wore a headmounted Sennheiser HMD-410 close-talking dynamic microphone and recorded their speech directly to DAT (digital audio tape) at a sampling rate of 48kHz. The monaural recordings were subsequently combined to facilitate two-channel stereo listening. The speakers did not see their partners or socialise with them outside of the recording sessions. Partner combinations were controlled for sex, age, and familiarity, and all recordings were transcribed and time-aligned for subsequent analysis.

In all, ten people took part in these recordings, five male and five female. Six were Japanese, two Chinese, and two native speakers of American or Australian English. All conversations were held in Japanese. The non-native speakers were living and working in Japan, competent in Japanese, but not at a level approaching native-speaker fluency. Partners were initially strangers to each other, but became friends over the period of the recordings. There were no constraints on the content of the conversations other than that they should last for thirty-minutes. Recordings continued for a maximum of ten sessions between native-speakers, and five with the non-native speakers. The speech data were transferred to a computer and transcribed manually using Wavesurfer public-domain speech transcription software [6] to provide a time-aligned record of what was spoken when, by who and to whom. The transcribed utterances were aligned to acoustic events in the speech waveform.

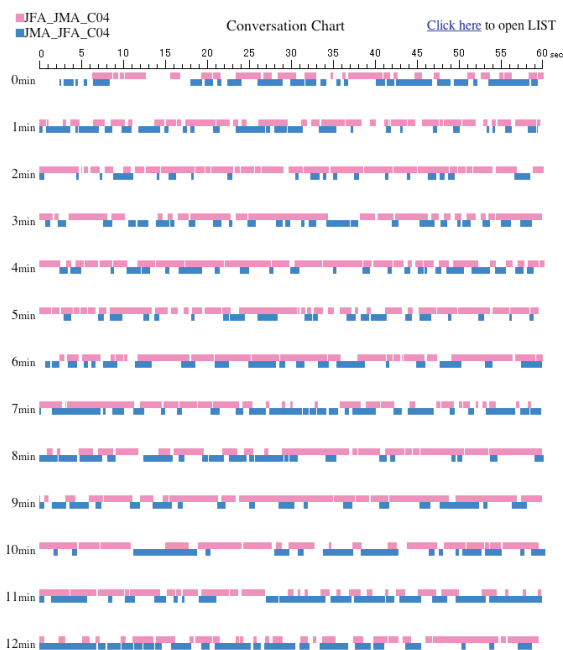


Figure 1: Showing speech activity in the first 12 minutes of the fourth conversation between a female and a male speaker. The patterns of long and short utterances clearly reveal the changes in dominance and the types of feedback activity in the conversation. (From <http://www.speech-data.jp/ta/esp-c/>).

### 1.3. Utterance-sized Speech Units

The definition of an ‘utterance’ in conversational speech is actually highly problematic. A common practice is to use any pause in the speech of greater than e.g., 200 milliseconds as an objective delimiting boundary, but it was noticed that even many single words contained ‘pauses’ of more than 300 milliseconds in these conversational data. It was therefore proposed that our transcribers should use a perception-based ‘one-yen-per-item’ criterion for segmenting the speech, whereby they would increase their payment for more items produced by cutting the speech into shorter utterances, but would be penalised for breaking up a single utterance into too small or ‘unnatural’ units.

The resulting segmentation was thus largely performed at the level of the phrase, or ‘minor intonation unit’, i.e., an utterance being a word or group of words demarcated by a single intonation contour. However, in many cases the transcribers actually produced longer and more complex utterance units, with some including punctuation marks such as commas, perhaps because of uncertainty about whether a clearly distinguishing intonational break could be heard.

Figure 1 shows an example of speech activity from the transcribed conversations such as can be found at the page [http://www.speech-data.jp/ta/esp-c/top\\_esp-c.html](http://www.speech-data.jp/ta/esp-c/top_esp-c.html), where the speech and text can be browsed interactively with a mouse. This type of coloured-bar plot reveals much information about the structure and dynamics of the conversation without requiring any indication of what was being said. Long continuous stretches of speech activity most likely indicate parts of the conversation with high propositional content; while the short bursts of overlapping activity probably indicate backchannel utterances showing agreement or interest. What is not so clear from plots such as Figure 1 is the gradual change in utterance length as speakers approach a turn-taking boundary.

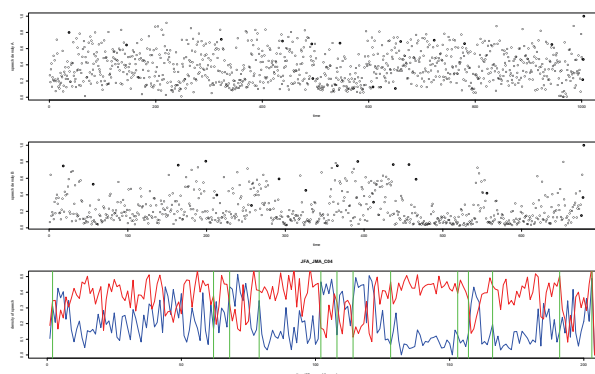


Figure 2: Speech activity plotted as the ratio of speech to silence per utterance for two speakers in a telephone conversation; a high value here represents almost continuous speech activity, whereas a low value indicates a brief utterance, typically giving feedback, in a long period of non-speech. The upper plot shows speaker A (1004 utts), and the middle plot speaker B (679 utts) with the same time axis of 30-minutes. The difference in the number of utterances is clear, as is the complementarity of the distributions. The lower plot shows the windowed and 10-second time-aligned values for both speakers superimposed.

## 2. Active Listening

Speakers take turns when engaging in a conversation, but this turn-taking activity appears from our analysis to be more complicated than previously envisaged. An examination of the telephone-speech data shows a high degree of complementary timing throughout each turn, resulting in a precise alignment of utterance durations and a compensation in length of backchannel or feedback utterances in accordance with the changing length of the partner’s utterance durations.

To examine the timing details of these spoken interactions, we produced a new measure of *utterance density*, or relative speech activity per unit of time by dividing each utterance duration by the sum of previous and following pause durations. Figure 2 shows a plot of this ‘utterance density’ for one of the 30-minute conversations. The figure shows density of speech activity against time. It is clear of course that the number of utterances differs between speakers but note also that the density of short utterances appears complementary as each provides feedback in turn to the talk of the other.

Since the number of utterances each partner produces per unit of time is of course highly variable, as is the length of each utterance, a direct comparison across the two speakers’ data is difficult. However, by summing and averaging the ‘density’ measures across each 10 second period of the conversation we produced a smoothed and time-aligned indication that allows a closer examination of the co-related speech activity. This is shown in the lower part of Figure 2, with data for both speakers plotted together in the same frame.

Figure 3 shows a detail taken from the first part of this plot. It is clear from this detail that not only do speakers take turns in dominating the conversation, but also that their precise timing of utterances within each turn follows a constantly changing and complementary pattern. We employed an Echo State Network (ESN) to model these fine changes in utterance timing and tested the result as a method of measuring ‘naturalness’ of a conversation.

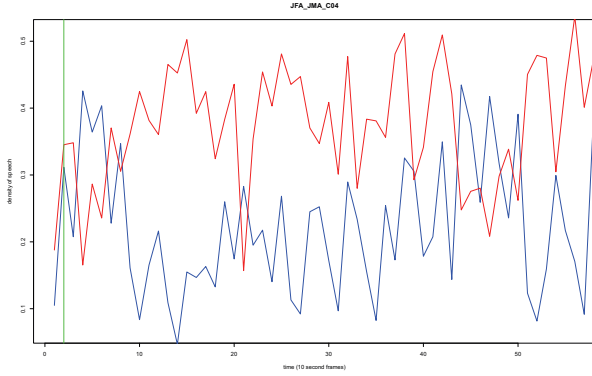


Figure 3: Showing convergence of speech activity at change of dominant speaker. This detail from the first ten minutes of the lower plot of Figure 2 (as also illustrated by the speech-bars in Figure 1) illustrates the type of gradual reciprocity we commonly found in the timing of utterances at the change of speaker. The red line is rising to point 15, while the blue falls over the same range, then red falls to point 22 while blue rises. After a brief crossover, red falls from 25 to 45 (while blue rises) then rises again from 45 while blue falls. This is NOT a simple switching of dominant speaker, but a synchronous interaction.

### 3. ESN Architecture and Training

In order to be able to predict the behaviour of the interlocutor using neural networks, recurrent neural networks (RNN), which take preceding observations into account for their predictions, are recommended. In the present paper a relatively novel kind of RNNs is used, the so called Echo State Network (ESN) [7]. Among the advantages of ESN learning over common RNN learning methods are the stability against noisy inputs [8] and the efficient direct pseudo inverse adaptation method to adapt the weights of the network [9].

As seen in Figure 4, the input layer  $K$  with its  $|K|$  neurons is fully connected to the dynamic reservoir  $M$  with  $|M|$  neurons, and the reservoir again is fully connected to the output layer  $L$  with  $|L|$  neurons. The fundamental part of the network is the so called reservoir, which is responsible for the dynamic characteristics, that are essential to be able to model a dynamic interlocutor behaviour. Typically, it is a collection of neurons, that are loosely connected to each other. The probability  $\beta$  of a connection  $w_{ij}$  between neuron  $m_i$  and neuron  $m_j$  to be set (i.e.  $w_{ij} \neq 0$ ) in the connection matrix  $W$  is set to  $\beta = 2 - 10\%$ , whereas the connections between the in- and output layer with the reservoir are all set. Since, there are feedback and recursive connections within the reservoir, not only the input is taken into account for the output but also the current state of each of the neurons and the history of all the preceding inputs, renders ESNs an ideal candidate for encoding dynamic processes such as movement patterns or non-verbal utterances [8, 10, 11].

The connection matrix is normalised to a so called spectral width parameter  $\alpha$  guaranteeing that the activity within the dynamic reservoir is kept at a certain level. In order to train an ESN it is only necessary to adapt the output weights  $W^{out}$  using the direct pseudo inverse method computing the optimal values for the weights from  $M$  to  $L$  by solving the linear equation system  $W^{out} = (S^+ T)^t$ , where  $t$  indicates the transpose operation and  $S^+$  indicates the pseudo inverse of the state collection matrix  $S$ .

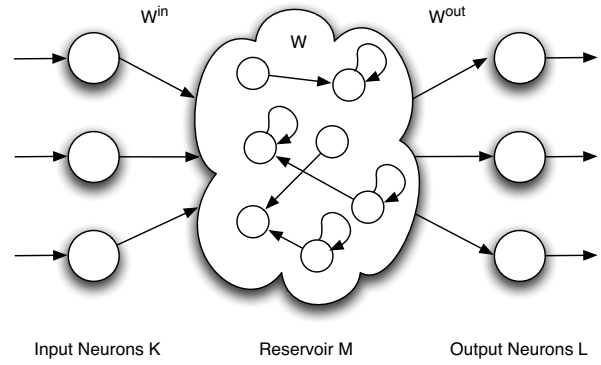


Figure 4: Scheme of an Echo state network. The input layer  $K$  consists of  $|K|$  neurons and is fully connected to the reservoir  $M$  with  $|M|$  neurons via the weights stored in the matrix  $W^{in}$ . Within the reservoir connections are set randomly and sparsely between the neurons of reservoir  $M$ . Entries  $w_{ij} \neq 0$  in the weight matrix  $W$  correspond to connections that are set between neuron  $m_i$  and  $m_j$ . The output layer  $L$  with  $|L|$  neurons is again fully connected to the reservoir via the weight matrix  $W^{out}$  that is adapted using the pseudo inverse method during the training of the network.

The method minimises the distance between the predicted output of the ESN and the target signal  $T$ : A network with  $|K|$  inputs,  $|M|$  internal neurons and  $|L|$  output neurons is considered as shown in Figure 4. Activations of input neurons at time step  $n$  are  $U(n) = (u_1(n), \dots, u_{|K|}(n))^t$ , of internal units are  $X(n) = (x_1(n), \dots, x_{|M|}(n))^t$ , and of output neurons are  $Y(n) = (y_1(n), \dots, y_{|L|}(n))^t$ . Weights for the input connection in an  $(|M| \times |K|)$  matrix are  $W^{in} = (w_{ij}^{in})$ , for the internal connection in an  $(|M| \times |M|)$  matrix are  $W = (w_{ij})$ , and for the connection to the output neurons in an  $(|L| \times |M|)$  matrix are  $W^{out} = (w_{ij}^{out})$ .

The activation of internal and output units is updated according to:

$$X(n+1) = f(W^{in}U(n+1) + WX(n)),$$

where  $f = (f_1, \dots, f_{|M|})$  are the internal neurons sigmoid transfer functions. The outputs are computed according to:

$$Y(n+1) = f^{out}(W^{out}X(n+1)),$$

where  $f^{out} = (f_1^{out}, \dots, f_{|L|}^{out})$  are the output neurons sigmoid transfer functions.

A detailed description of the offline learning procedure is given in the following:

1. Given I/O training sequence  $(U(n), D(n))$
2. Generate randomly the matrices  $(W^{in}, W, W^{back})$ , scaling the weight matrix  $W$  such that it's maximum eigenvalue  $|\lambda_{max}| \leq 1$ .
3. Drive the network using the training I/O training data, by computing

$$X(n+1) = f(W^{in}U(n+1) + WX(n))$$

4. Collect at each time the state  $X(n)$  as a new row into a state collection matrix  $S$ , and collect similarly at each time the sigmoid-inverted teacher output  $\tanh^{-1} D(n)$  into a teacher collection matrix  $T$ .

5. Compute the pseudo inverse  $S^+$  of  $S$  and put

$$W^{out} = (S^+T)^t$$

If the output exceeds a threshold  $\theta$  it is counted as a hit and the output is set to 1 (the interlocutor is active), values below  $\theta$  are set to -1 (the interlocutor is not active).

## 4. Results

The ESN was trained using the following parameters:  $|M| = 100$ ,  $|K| = 30$ ,  $|L| = 1$ ,  $\alpha = 0.05$ ,  $\beta = 2\%$ ,  $\theta = 0$ , and  $f(x) = \tanh(x)$ . Input signal  $U$  and target signal  $D$  were taken from the annotations of the telephone conversations.

The ESN was tested using a subset of the corpus comprising of twenty conversations from speaker JFA talking to a male and female partner respectively. The timings of utterance onsets and their durations were transformed into signals with a sampling rate of 10 Hz, with the data of the interlocutor used as input and that of JFA as the target  $\in \{-1, 1\}$ . Density values were predicted from the framewise estimates.

We provided context information to the ESN by folding the one dimensional input signal  $U$  into a 30 dimensional signal (also sampled with a frequency of 10 Hz) i.e., the one dimensional input over the last 3 seconds is transposed and forms one single observation of the new signal. Each observation is followed by another 30 dimensional observation which contains the next element of the one dimensional signal and the rest of a 3 second window. The ESN was tested using a cross-validation leave-one-out paradigm, trained on 9 conversations per partner and tested on the remaining one. The network predicted speech density values for JFA from each frame of her partner's speech.

Table 1 shows the error in predicted speech-density measures compared with observed values. It is clear from the table that the ESN performs significantly better than random and that it predicts more natural interaction sequences than found when switching partner timings from other natural conversations.

The ESN proved capable of modelling the speaking behaviour of JFA over different conversation partners. From the small size of the error we can infer that it was accurately modelling not just the change of speaker, but also the changes in utterance duration throughout each turn.

Future work will focus more closely on these changes in the synchrony of a mutual discourse and will attempt to explain the mechanism whereby speakers are able to adjust their performance so closely.

Table 1: *Echo State Network Prediction Errors. 'Conv' shows the conversation number for speaker JFA with a male and female interlocutor. M9 and F9 show ESN cross-validation results. M-F and F-M show results when trained on male and tested with female data and vice versa. RND shows the result with randomised responses. M-sw and F-sw show non-random results for switching the male and female partner responses:*

Conv	1	2	3	4	5	6	7	8	9	10
M9	.11	.13	.13	.11	.14	.15	.13	.13	.17	.15
F9	.12	.11	.12	.11	.11	.11	.12	.12	.12	.14
M-F	.11	.11	.12	.12	.13	.11	.13	.13	.15	.15
F-M	.13	.13	.14	.13	.16	.16	.14	.15	.15	.17
RND	.29	.31	.28	.31	.28	.26	.31	.31	.28	.26
M-sw	.24	.23	.24	.23	.25	.25	.24	.23	.27	.27
F-sw	.24	.25	.24	.24	.22	.23	.25	.24	.26	.28

## 5. Conclusion

This paper has presented the results of an analysis of timing details of conversational speech and has shown that participants fine-tune their utterance lengths in a precise way to facilitate turn-taking. We attempted to model this behaviour for an automated conversation system using an echo-state network and were able to predict the timing activity at levels significantly better than chance. These findings confirm Kendon's view of frame attunement in discourse.

## 6. Acknowledgements

This work was carried out at Trinity College Dublin while the second author was visiting under a Companion Technology Project (SFB-TRR-62) funded by the German Research Foundation. We are grateful to the JST for funding the creation of the Expressive Speech Processing corpus.

## 7. References

- [1] Hutchby, Ian, (1999) "Frame attunement and footing in the organisation of talk radio openings", *Journal of Sociolinguistics* 3/1, pp.41-63.
- [2] Goffman, Erving. 1974. *Frame Analysis*. New York: Harper and Row.
- [3] Kendon, Adam, (1990) *Conducting Interaction: Patterns of Behaviour in Focused Encounters*. Cambridge: Cambridge University Press.
- [4] Campbell, N., "Individual Traits of Speaking Style and Speech Rhythm in a Spoken Discourse", pp.106-120 in A. Esposito et al. (Eds.): *HH and HM Interaction 2007*, LNAI 5042, pp. 107120, 2008, Springer-Verlag Berlin Heidelberg 2008
- [5] Campbell, N., "An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal Conversation Data", (Wed-Ses2-S1), in *Proc Interspeech 2009*, Brighton.
- [6] Campbell, N., "Speech & Expression; the Value of a Longitudinal Corpus", in *Proc 4th International Conference on Language Resources and Evaluation*, pp183-186, 2004
- [7] Jäger, H. and Haas, H.: *Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication*, *Science*, pp. 78-80, Vol. 304, 2004.
- [8] Scherer, S. and Oubbati, M. and Schwenker, F. and Palm, G.: *Real-Time Emotion Recognition from Speech Using Echo State Networks*, in *Proc. of the 3rd IAPR workshop on Artificial Neural Networks in Pattern Recognition (ANNPR08)*, pp. 205-216, 2008.
- [9] Jaeger, H.: *Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach*, Tech. Report Fraunhofer-Gesellschaft, St. Augustin Germany, Nr. 159, 2002.
- [10] Scherer, S. and Schwenker, F. and Campbell, W. N. and Palm, G.: *Multimodal laughter detection in natural discourses*, in *Proc. of 3rd International Workshop on Human-Centered Robotic Systems (HCRS'09)*, pp. 111-121, 2009.
- [11] Krause, A. F. and Blaesing, B. and Duerr, V. and Schack, T.: *Direct Control of an Active Tactile Sensor Using Echo State Networks*, in *Proc HCRS'09*, pp. 11-21, 2009.