2009-01-01

Convergence in Human Dialogues Time Series Analysis of Acoustic Feature

Spyros Kousidis
Technological University Dublin, spyros.kousidis@tudublin.ie

David Dorran
Technological University Dublin, david.dorran@tudublin.ie

Ciaran Mcdonnell
Technological University Dublin, cmcdonnell@tudublin.ie

See next page for additional authors

Recommended Citation

Kousidis, S. et al. (2009) Convergence in Human Dialogues Time Series Analysis of Acoustic Feature. *SPECOM 2009,* St. Petersburg, Russian Federation.

Authors

Spyros Kousidis, David Dorran, Ciaran Mcdonnell, and Eugene Coyle

Time Series Analysis of Acoustic Feature Convergence in Human Dialogues

Spyros Kousidis[1], *David Dorran*[2], *Ciaran McDonnell*[1] *& Eugene Coyle*[2]

[1]Digital Media Center; [2]Audio Research Group;
[1,2]Dublin Institute of Technology, Ireland
spyros.kousidis@dit.ie

Abstract

Convergence of acoustic/prosodic (a/p) features between two speakers is a well-known property of human dialogue. It has been suggested that this particular aspect of human interaction should be implemented in spoken dialogue systems, so that they can be perceived as more "human-like". This paper presents a quantitative analysis method that can provide information required for modeling the phenomenon of convergence. The analysis is a combination of TAMA, a previously introduced data extraction method, and bivariate time series analysis. Results show significant correlation of a/p features between speaker dyads in the recorded dialogues analyzed, and indicate a significant amount of feedback, which a statistical verification of *bi-directional* convergence.

1. Introduction

Current advances in spoken dialogue systems [1, 2] point towards a direction of more "human-like" interfaces. For certain applications, it is desired that users can perceive a system through the *human metaphor*, i.e. as if they were talking to a human being, rather than a machine. This is pursued by identifying properties of human dialogue speech and building them into the system. One such property is convergence of acoustic/prosodic (a/p) features between two (human) speakers. Therefore, a practical model for convergence would provide the means to realize a more realistic "human-like" behaviour.

The approach presented in this paper describes the phenomenon in a quantitative way, by combining the TAMA method [3] with time-series analysis.

1.1. Convergence

Convergence is defined as a situation where "the observed behaviors of two interactants, although dissimilar at the start of the interaction, are moving towards behavioral matching" [4]. In plain terms, convergence refers to speakers' adaptation of their interactive behaviour (including properties of speech) to that of their dialogue partners.

The phenomenon has been studied in various fields of research, including psycholinguistics, behavioral sciences, and communication science. The majority of these studies attribute one or more functions to convergence. There is a range of such functions, from autonomous, non-intended behaviour to habitual behaviour and even to intentional communicative strategies [4-8]. However, there is general agreement that convergence is a sign of positive evaluation towards the partner and that it is also positively evaluated by the partner [7]. Therefore, spoken dialogue systems that can simulate this behaviour are likely to be more "appealing" to the user.

Convergence between two speakers A and B can be both unidirectional (A → B) as well as bidirectional (A ↔ B). Also, convergence can be unimodal or multimodal. This refers to the number – one or many – of different dimensions (properties of speech communication) along which the speakers can converge simultaneously [6]. Such dimensions can be the choice of words, syntax, pronunciation, regional/ethnic accent, tone, rhythm, loudness, facial gestures and body posture. The analysis presented here focuses on convergence of a/p features, namely pitch, intensity, speech rate, and pitch range.

1.2. Convergence in human-machine interaction

Studies in human computer interaction have shown that human users adapt linguistically to interfaces even when using only text input [9]. For speech, it has been reported in [10] and [11] that users were found to adapt to prosodic and temporal characteristics of a 'talking' system. This has been utilized in [12], where the users unknowingly adapted their speech rate to that of the spoken dialogue system, which was designed to "keep" their speech rate within limits where automatic speech recognition (ASR) performance was higher. Further, it was reported in [13] that users showed preference towards an interactive voice response system that adapted its own speech rate according to their own, which is a strong indication that convergence of machines towards users is positively evaluated.

Therefore, convergence can already be utilized to make dialogue systems more appealing, as well as improve their performance. However, a well-developed quantitative model for convergence does not exist yet. Such a model is essential for unlocking the full potential of utilizing convergence in order to design more "human-like" dialogue systems.

1.3. Towards "human-like" convergence

Following the evaluation framework described in [1], implementing convergence in spoken dialogue systems requires prior knowledge of the process in human dialogues. By definition (see section 1.1), convergence *is a continuous and bidirectional process that evolves over time*. These properties of the natural process, point directly towards time series analysis. This holds true, whether the objective of the analysis is a description, a model, forecasting, or the application of *monitoring and control* on the process [14].

The mode of convergence depends heavily on speaker personality, gender, and context (application). Therefore, a spoken dialogue system will have to adapt its control strategy on-line, according to observed user behaviour. As a result, the methodology presented here focuses on unsupervised methods for extracting a/p features and other information from the audio signal.

2. Speech corpus acquisition

Dialogues between adult native English speakers were recorded for this study. The subjects were communicating through microphones and headphones, while sitting in soundproof isolation booths equipped with monitors. There was no visual contact between the interactants during the dialogues.

2.1. Experimental scenarios

Three experimental scenarios were presented, in which the subjects were required to (verbally) cooperate in order to survive in a hypothetical adventurous situation (see Figure 1). A collection of 15 items (identical for both subjects) was displayed on the monitors, and the subjects were asked to freely discuss and reach an agreement on the order of importance; the item considered to be the most important and essential to survive the hypothetical hazard was given rank 1, the next most important was given rank2, and so on until all 15 items had been ranked.

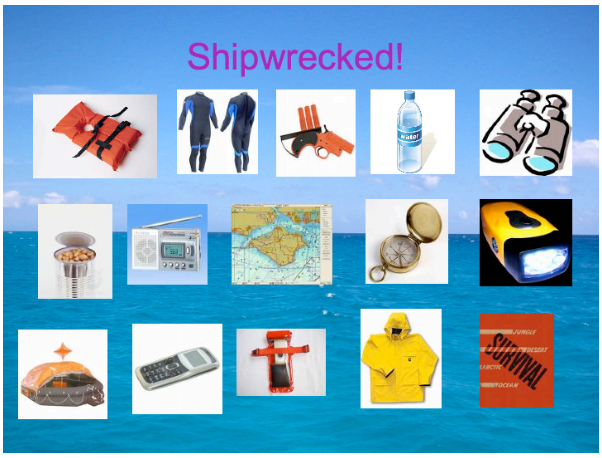


Figure 1- "shipwrecked" scenario

The three scenarios involved the two subjects being shipwrecked, stranded in space in a pod, or lost from their group in the snowy Himalayas. In all three situations, the subjects had to "survive" long enough until rescued (a time limit of 10 minutes applied). In addition, the subjects were provided only with pictures of the items and were instructed to decide themselves on the name/description of an unknown item.

Due to the low difficulty of the task and the general lack of constraints in the experimental design, the speech corpus contains a substantial share of spontaneous speech and dialogue acts, a significant amount of laughter and other non-speech elements, and many occurrences of overlapping speech.

The entire sessions were recorded in separate audio channels for each speaker (from the microphones in each booth), with very high audio quality (192KHz/24-bit).

2.2. Segmentation and feature extraction

The recorded files were down-sampled to 44.1 KHz/16-bit prior to analysis with the freely available software Praat [15]. For each audio file (that contains the entire speech stream from one of the two speakers), the following actions were performed: first, pauses were detected by use of an intensity and duration threshold. This process is automatic, but manual corrections were required in order to eliminate noise classified as speech. In addition, non-speech elements such as laughter and breath noises were manually annotated.

The two resulting timelines (one for each speaker) contain marked boundaries for speech, non-speech elements, and pauses. Combined, they produce a *chronograph* [16] of the dialogue, i.e. a schematic of turn switching and overlap between the two speakers (see Figure 2).

The a/p features extracted for each marked speech interval were average pitch, pitch range, average intensity and vowel detection. The latter is used as an estimate of speech rate (number of vowels per minute) [17]. Pitch range is defined here as equal to two standard deviations (computed together with average pitch).

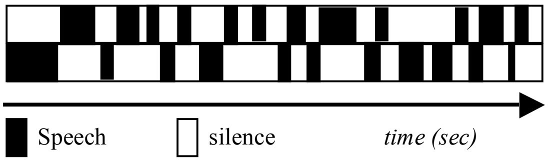


Figure 2 - Chronograph of part of dialogue

2.3. The TAMA method

Points for the time-series analysis were acquired by implementation of the TAMA method, which calculates average values of a/p features for a series of overlapping frames of fixed length (see Figure 3). The process is equivalent to a *simple moving average* filter, hence the name (time-aligned moving average).

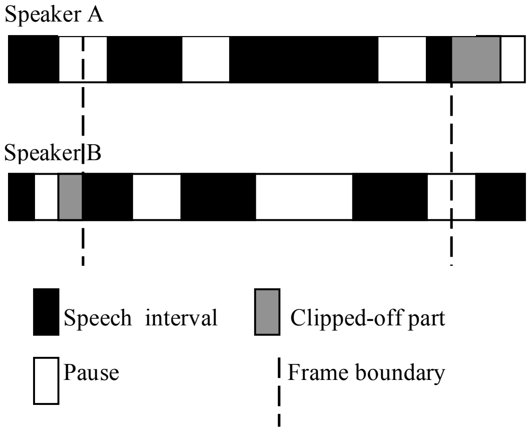


Figure 3 - Schematic of TAMA frame

The average value of each feature is calculated using equation (1) below:

$$\mu = \sum_{i=1}^{N} f_i d_i \Big/ \sum_{i=1}^{N} d_i \qquad (1)$$

where μ is the average, i is the interval index, N is the total number of intervals, f_i is the value of the feature for interval i, and d_i is the duration of the interval i.

If the average of a feature for the entire dialogue is being calculated, then i runs through all the speech intervals. If a frame average is being calculated, then i runs through all intervals that exist in the frame. For intervals that cross frame boundaries, the value of d_i is set to the duration of the interval *within* the frame. Equation (1) is essentially a *weighted mean*, where the interval durations d_i are the (un-normalized) weights. The normalized weights are defined as $w_i = d_i / D$, where $D = \Sigma d_i$, with $\Sigma w_i = 1$, in which case the standard error (S.E.) is given by

$$S.E. = \sqrt{\sum_{i=1}^{N} w_i^2 \sigma_i^2} \qquad (2)$$

where σ_i is the standard deviation of feature f_i (obviously not defined for pitch range and number of vowels).

The resulting averages of the frames are divided by the speaker's overall average (for the whole dialogue), giving a "normalized" feature value. This is deemed essential in order to make meaningful comparisons between speakers with largely different inherent speech characteristics (such as male vs female speakers). Effectively, the normalization changes the random variable from an a/p feature to a *dimensionless variable* with mean equal to one. The deviation of each point from the mean is equivalent to a proportional increase or decrease relative to the mean, e.g. a value of 1.2 for pitch represents a frame were the average pitch is 20 percent higher than the overall average pitch of the speaker in the dialogue..

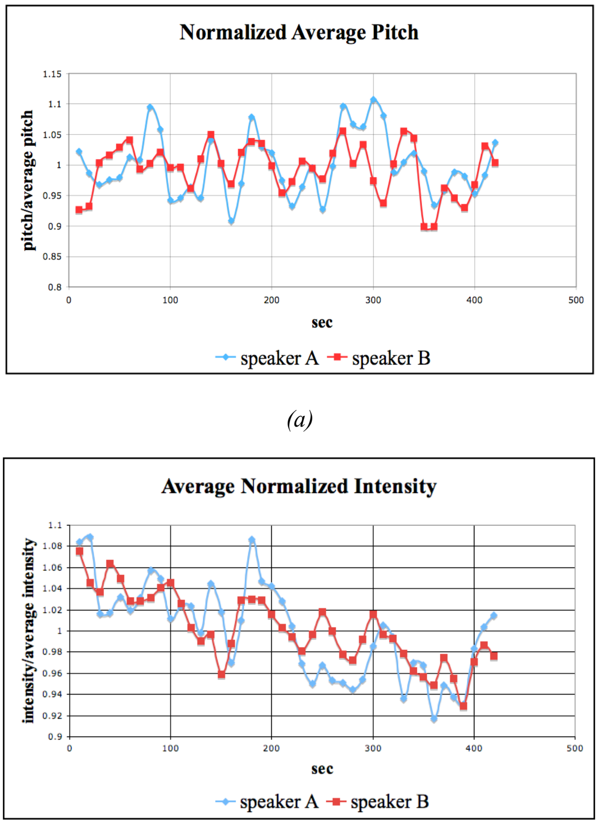
2.4. Time series analysis

As suggested in many statistics textbooks on time series analysis (for example [14]), the first step in the analysis of a process is to illustrate its *time plot* (see Figure 4*(a)* and 4*(b)*).A TAMA frame with a length of 20 seconds and a time step of 10 seconds has been used, in order to provide the data points for these plots. This results in a 50% overlap (the first half of each TAMA frame is the second half of previous frame), thus the points are equally spaced at ten seconds apart.

The process of frame length selection is equivalent to applying an appropriate moving average filter to a time series in order to reduce the variance and provide a smoother curve, where trends or other features may be more easily identifiable [3]. There are two constraints that can be helpful to estimate an appropriate frame length. First, a short length may yield empty frames (where one speaker holds the turn and the other is silent), resulting in a time series with missing values. Although these can be dealt with, either by using the most recent value (from previous frame) or by linear interpolation, it is desirable to avoid them. The second constraint is that overly large frame lengths tend to "over-smooth" the series.

The plot in Figure 4*(a)* indicates that the two series are converging to a certain degree, as expected. Also, each series (individually) appears to have an *autoregressive* structure, as it can be seen from the plot that consecutive values are likely to be close to each other. This indicates that a/p features of speakers change smoothly over time, unless an event occurs (such as a topic change) that changes the flow of the dialogue.

However, if convergence occurs, any value of series A will also depend on contemporaneous and previous values of series B, and vice versa. In other words, *feedback* is expected to occur between the two series.



(a)

(b)

Figure 4 - Time plots of (a) pitch and (b) intensity for two speakers (A,B). Feature averages of 20 second frames with 50% overlap (normalized values)

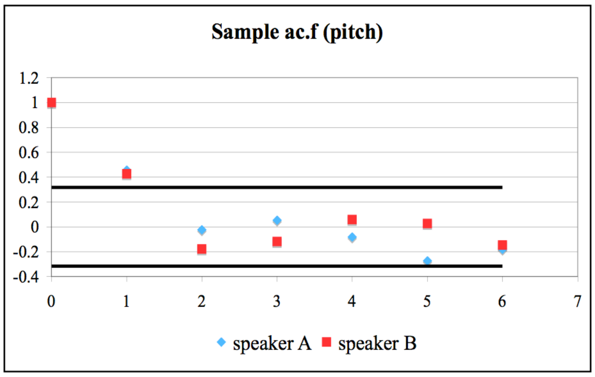


Figure 5 - Correlograms of the two individual series (speakers A and B) shown in Figure 4a

Information for each individual series can be extracted by use of the *sample autocorrelation function (ac.f)*. A good estimate of the ac.f is the *correlogram* (see Figure 5). The sample autocorrelation coefficient r_k at lag k is given by

$$r_k = \frac{\sum_{t=k+1}^{n} (x_t - \mu)(x_{t-k} - \mu)}{\sum_{t=1}^{n} (x_t - \mu)^2} \tag{3}$$

where x_t are the series values and μ is the sample mean calculated with equation (1).

The correlograms of both individual series in Figure 5 quickly drop to zero (with 95% confidence limits at ±2/√n), an indication that the processes are *stationary.* This is not always the case; in Figure 4*(b),* the plot of average intensity shows a global decreasing trend for both speakers. The correlogram in Figure 6 shows that the two series are not stationary. Stationarity can be achieved by *differencing*, i.e. subtracting the previous value from the current value in the series. The process can be repeated several times, until the resulting series is stationary. If d repetitions are required, the series is said to be *integrated of order d,* and denoted by ***I(d)***. The significance of this is explained in section 2.5.

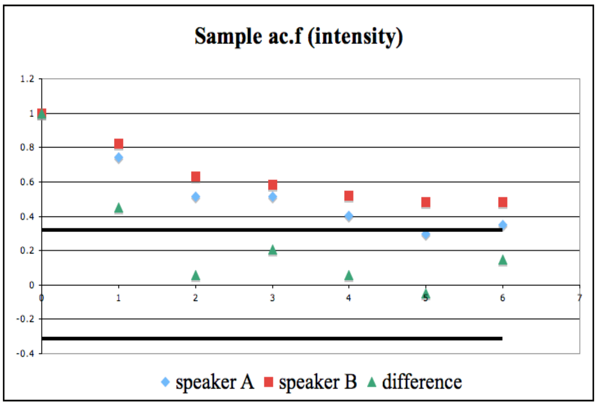


Figure 6 - Correlograms of the two individual series shown in Figure 4b (speakers A and B), and their difference, (A-B)

2.5. Bivariate time series

In order to evaluate whether two time series are causally related, one has to turn to bivariate time series analysis. This type of analysis considers two time series as components of a *linear system,* where one of the series can be regarded as the *input* and the other as the *output.* However, if the input series is also affected by the output series, then *feedback* is present. In such cases, the results have to be considered carefully, as they may be misleading.

The relationship between two series can be explored by use of the *cross-correlation function* (cc.f), which measures dependence between values of one series to past values of the *other* series. The cc.f can be estimated by use of the *cross-correlogram*, which is a plot of cross-correlation coefficients at different lags (Figure 7).

The sample cross-correlation coefficient $r_{xy}(k)$ at lag k is given by

$$r_{xy}(k) = \begin{cases} \dfrac{\sum_{t=k+1}^{n} (x_t - \mu_x)(y_{t-k} - \mu_y)}{\sqrt{\sum_{t=1}^{n} (x_t - \mu_x)^2 \sum_{t=1}^{n} (y_t - \mu_y)^2}}, k \geq 0 \\ \dfrac{\sum_{t=k+1}^{n} (x_t - \mu_x)(y_{t-k} - \mu_y)}{\sqrt{\sum_{t=1}^{n} (x_t - \mu_x)^2 \sum_{t=1}^{n} (y_t - \mu_y)^2}}, k < 0 \end{cases} , \tag{4}$$

where x_t , y_t are the values of the two series, and μ_x , μ_y are the overall averages for those series, calculated by equation (1).

As suggested in [14], spuriously large coefficients may appear in the cross-correlogram if the two individual series have not been previously "pre-whitened", i.e. converted into white noise. This can be achieved by fitting an appropriate model to each series. In the correlogram of Figure 5, for example, both series of average normalized pitch have only one significant coefficient at lag 1, with a value around *0.4*. This can be used as an *alpha* value for an *autoregressive* (AR) model of order 1, which is fitted to each series. The resulting *residual* series are then tested for cross-correlation (Figure 7).

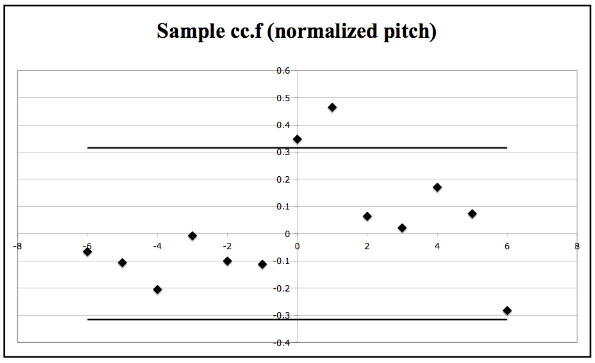


Figure 7 - Sample cross-correlation (pitch) between two (speakers A, B). (95% confidence limits at ±2/√n)

There two positive coefficients that are significantly different from zero, at lags 0 and 1. A large coefficient at lag zero is an indication of the presence of feedback. In the context of convergence analysis, feedback is a result of *both* speakers converging towards each other (bi-directional convergence). Positive or negative lags represent unidirectional convergence (A→B and B→A respectively). However, in the presence of feedback, interpretation of the cross-correlogram can be misleading (see discussion). The bivariate process can be described by a *vector autoregressive* (VAR) model of the form

$$X_t = \Phi X_{t-1} + \varepsilon_t ,$$
$$X_t = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \varepsilon_t = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}, \Phi = \begin{pmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \phi_{22} \end{pmatrix}, \tag{5}$$

where x_1,x_2 are the values of the two series and ε_t is the *error vector.*

Dialog		AvP (Hz)	AvI (dB)	PR (Hz)	SR (v/m)	TFS (sec)	TTS (%)	TO (%)	TP (%)	TDD (sec)	Significant coefficients (lags)			
											AvP	AvI	PR	SR
1	F	191	54	64	241	74	20.5	17.2	18.3	428	0,1	0,1	1	-1
	M	138	59	46	173	181	43.9							
2	M	117	62	27	226	169	35.5	15.4	16.31	492	0	0	0	-
	M	146	61	49	178	146	32.9							
3	M	141	54	51	210	145	38.4	44.5	23.5	390	1	0	-	1
	F	222	54	120	192	99	26.7							
4	F	203	56	128	240	124	28.7	11.9	37.9	516.9	0	0	0	-
	F	212	54	90	225	103	21.5							
5	M	140	63	41	183	125	36.3	50	19.1	360.3	0	0	-	0
	M	163	52	57	177	108	30.7							

Table 1- Results of feature extraction and time series analysis. Average pitch (AvP), average intensity (AvI), speech rate (SR), pitch range (PR), total fluent speech (TFS), total turn share (TTS), total overlap (TO), total pauses(TP) and total dialogue duration (TDD) are shown.(F) or (M) denote a female or a male speaker, respectively. The numbers at the four rightmost columns indicate the lags at which a significant positive correlation is found.

If the *parameter matrix* $\boldsymbol{\Phi}$ is triangular, then the system is said to be *open-loop* and can be modeled by a linear equation between the two variables. If both $\boldsymbol{\varphi_{12}}$ and $\boldsymbol{\varphi_{21}}$ are large, then the system demonstrates *feedback* and is said to be *closed-loop.*

The estimation of matrix $\boldsymbol{\Phi}$ can be rather complex if feedback is present. The case in Figure 6 is yet more complicated: the individual series are non-stationary and *I(1)*, but their difference $(\boldsymbol{x_1 - x_2})$, which is a linear combination of the two variables, *is I(0)* and stationary. Therefore, the two series are *co-integrated*, a fact that can perhaps be helpful in identifying a more appropriate model, that includes the *co-integration vector* $a^T = (1,-1)$. Appropriate models for convergence will be considered in the future.

3. Results and discussion

Results are shown in Table 1. Positive correlations at lag zero were expected, as they are the result of bi-directional convergence. In the five different dialogues analyzed here, significant coefficients at lag zero were found for all features studied. This was more evident for pitch and intensity (a large coefficient at lag zero was found for all dialogues studied). Similar, but less conclusive results were found for average pitch range and speech rate (significant coefficients in some of the dialogues). Significant coefficients were also found at lags -1 and 1. Theoretically, this indicates unidirectional convergence, either from (speaker) A to B or vice versa, with a lag of 10 seconds (determined by the chosen frame length). However, such an interpretation would be naïve, especially in case a large coefficient at lag zero is also present. As a result of the TAMA process, some of the autoregressive properties of convergence are "included" in lag zero, as it represents a time frame. At best, the presence of large coefficients at lags -1 or 1 can only be interpreted as an indication that convergence *does* have autoregressive properties.

The importance of these results is that convergence of a/p features over time can be statistically evaluated. Although the cross-correlation analysis cannot be directly relied upon for parameter estimation, it does so for *model identification.* The autoregressive structure of the individual series points towards a VAR model with large *feedback terms.* In addition, the co-integration vector (1, -1) suggests that the *difference,* $(x_1 - x_2)$, or in other words the *distance* between the speakers is important. These are only initial hypotheses and further work is required before a model can be formulated. Assuming that a VAR or similar model (from the VARMAX family) is the most appropriate, the summary statistics can be used to design adaptive control of convergence in a dialogue system. Systems of this type will be able to employ "strategies", such as leaving the initiative to the users and converge to their style, monitoring if the user is converging and taking action to encourage convergence, or a mixture of both. However, a single model/strategy is unlikely to be appropriate for an entire dialogue, because the a/p features of any given utterance are not a *function* of convergence, but rather of many exogenous factors (type of utterance, topic changes, errors). A combination with other dialogue monitoring functions of dialogue systems, such as state-space dialogue modeling (e.g. [18]) may allow deployment of more appropriate convergence strategies/models for different dialogue states.

The justification of the TAMA method, i.e. using frames rather than utterances as units, merits discussion. The convenience for analysis introduced by the transformation of the data is not negligible, (analysis of series with points at irregular intervals is less straightforward), but there are also additional advantages. In an (assumed) adaptive system that uses TAMA frames in order to monitor and control convergence, the frame length can be used as a trade-off variable. Longer frames ensure smooth changes in a/p features and more stability, but shorter frames enable quicker response to sudden changes. In an utterance-based system, this can only be achieved by taking several preceding utterances into the calculation, probably weighing them to promote more recent ones, as was done in [19]. However, due to the complex structure of spontaneous speech, the most recent utterance may not be the most relevant. This problem can be partially overcome by utterance/dialog act classification [20]. However, employment of such techniques in real-time environments may increase computational load and introduce latencies. TAMA is a more crude method, but is also less demanding in resources. In addition, the TAMA method is virtually independent from the ASR component (although it would be desirable for some stages of feature extraction to be shared for economy).

Furthermore, it should be considered whether analysis of spontaneous (or unconstrained) human dialogues is the best way to model convergence for dialogue systems. The latter, in their majority, have specific applications and limited responses, so perhaps application-specific training is more appropriate; but this point of view is not compatible to the human metaphor paradigm described in [1, 2]. According to the framework proposed in [1], a system can only be evaluated against a *human* dialogue. Besides, it is difficult to train a system on an application environment that is not yet developed. Wizard-of-oz scenarios can be employed to overcome this problem [10], but care should be taken that design constraints do not bias analysis of convergence. A combination of analyses on human dialogues and wizard-of-oz scenarios is worthy of investigation.

After a suitable model has been identified, “human-like” convergence will be feasible to implement in existing or newly developed dialogue system architectures. Such systems will be able to employ unsupervised adaptive control, in order to simultaneously *monitor* as well as *simulate* convergence along multiple possible dimensions (different a/p features). *Control theory* is key here, mainly because of its wide repertoire of techniques for dealing with closed-loop systems that demonstrate feedback.

4. Conclusions and further work

Time series analysis can provide useful insights into the process of a/p convergence in human dialogues. This will be useful in developing spoken dialogue systems that can display similar behaviour, in order to help the user visualize the human metaphor.

Accumulation of results from analysis of dialogue recordings is required before sufficient knowledge of the process is gained. In the future, different recording experiments, either of human dialogues, or wizard-of-oz scenarios, that simulate real application environments, will be considered.

The analysis method presented here is *feature independent*, and can thus be readily modified in order to analyze convergence of other properties of speech, e.g. pause duration, intonation, or even lexical and/or gesture features. This will enable investigation of redundancy among different “convergence channels”, by correlating several distinct features *simultaneously*. Careful consideration will then be required to deal with the complexity of the resulting *n-variate* time series analysis.

5. Acknowledgements

A substantial part of this research was funded by the European Union as part of the *SALERO* project (www.salero.info) through the *IST programme* under FP6. Part of the research was also funded by Dublin Institute of Technology (www.dit.ie). The authors would like to thank Dr Charlie Cullen and Brian Vaughan for their assistance in the experimental setup and recording of dialogues.

6. References

[1] J. Edlund, J. Gustafson, M. Heldnera, and A. Hjalmarssona, "Towards human-like spoken dialogue systems," *Speech communication,* vol. 50, pp. 630-645, 2008.

[2] R. Carlson, J. Edlund, M. Heldner, A. Hjalmarsson, D. House, and G. Skantze, "Towards human-like behaviour in spoken dialog systems," in *Swedish Language Technology Conference (SLTC)* Gothenburg, Sweden, 2006.

[3] S. Kousidis, D. Dorran, Y. Wang, B. Vaughan, C. Cullen, D. Campbell, c. McDonnell, and E. Coyle, "Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues " in *Interspeech 2008* Brisbane, Australia, 2008.

[4] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns*: Cambridge university Press, 1995.

[5] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences,* vol. 27, pp. 169-190, April 2004.

[6] H. Giles, A. Mulac, J. J. Bradac, and P. Johnson, "Speech Accomodation Theory: The First Decade and Beyond," in *Communication Yearbook 10*, M. L. McLaughlin, Ed. Newbury Park: SAGE, 1987, pp. 13-48.

[7] J. Welkowitz and M. Kuc, "Interrelationships Among Warmth, Genuineness, Empathy, And Temporal Speech Patterns In Interpersonal Interaction," *Journal of Consulting And Clinical Psychology,* vol. 41, p. 472, 1973.

[8] A. Ward and D. Litman, "Dialog Convergence and Learning," in *13th International Conference on Artificial Intelligence in Education*, Los Angeles, CA, 2007.

[9] E. Zoltan-Ford, "How to get people to say and type what computers can understand," *Int. J. Man-Mach. Stud.,* vol. 34, pp. 527-547, 1991.

[10] S. Oviatt, C. Darves, and R. Coulston, "Toward adaptive conversational interfaces: Modeling speech convergence with animated personas," *ACM Trans. Comput.-Hum. Interact.,* vol. 11, pp. 300-328, 2004.

[11] N. Suzuki and Y. Katagiri, "Prosodic alignment in human-computer interaction," *Connection Science,* vol. 19, pp. 131 - 141, 2007.

[12] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *ICPhS*, Barcelona, 2003, pp. 2453-2456.

[13] N. Ward and S. Nakagawa, "Automatic User-Adaptive Speaking Rate Selection," *International Journal of Speech Technology,* pp. 259-268, October 2004.

[14] C. Chatfield, *The Analysis of Time Series - An Introduction*, 5th ed.: Chapman & Hall/CRC, 1996.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 4.4.18 ed, 2006.

[16] M. Lennes and H. Anttila, "Prosodic features associated with the distribution of turns in Finnish informal dialogues," in *The Phonetics Symposium 2002. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing*, 2002, pp. 149-158.

[17] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing,* Seattle, WA, USA, 1998, pp. 945-948.

[18] K. Jokinen, "Natural Interaction in Spoken Dialogue Systems," in *Workshop Ontologies and Multilinguality in User Interfaces. HCI International* Crete Greece, June 2003, pp. 730-734.

[19] R. Nishimura, N. Kitaoka, and S. Nakagawa, "Analysis of Relationship Between Impression of Human-to-Human Conversations and Prosodic Change and Its Modeling," in *Interspeech* Brisbane, Australia, 2008.

[20] H. F. Wright, "Modelling prosodic and dialogue information for automatic speech recognition." vol. PhD: Unviresity of Edingburgh, 1999.