Prosodic features of simultaneous interpreting

George Christodoulides

george@mycontent.gr
Université de Louvain, Louvain-la-Neuve, Belgium

Abstract

We study the prosodic features of simultaneous conference interpreting in an attempt to describe its particular speaking style. We focus on the temporal organisation of the interpreters' speech (pauses, speech rate), as well as global prosodic properties (f_0 range, melodic agitation). We also study similarity and convergence phenomena between the speaker and the interpreter, on prosodic features, and their dynamic evolution over time. The findings indicate that interpreters make longer silent pauses, less frequently than speakers and their speech rate is more variable. In most cases, interpreters had a narrower pitch range than speakers and do not mirror the pitch of their speakers.

1. Introduction

Simultaneous interpreting (SI) has been described as a taxing cognitive task, during which the interpreter is working at the limits of their processing capacity (Pöchhacker 2004). Speaking style (*phonostyle*) is determined both by the situational context and by individual characteristics (Llisteri 1992; Eskenazi 1993; Léon 1993; Simon et al. 2010). The interpreter may choose to alter some local prosodic characteristics of their speech to mimic choices made by the speaker (Couper-Kuhlen & Selting 1996) if they deem it appropriate. The interplay of all these factors will define the speech style of the interpreter. Some interpreters may adopt a uniform, personal style regardless of the speaker, while others will be more influenced by and converge with the speaker.

This study focuses on conference interpreting of speeches delivered in public, in a political context. Our hypothesis is that the time and cognitive constraints of SI create a specific speaking style of interpreting. We also investigate whether and to what extent the prosodic characteristics of an interpreter's speech are influenced by those of the speaker and their evolution over time.

2. Corpus

2.1 Design and structure

A parallel bilingual spoken corpus was built for this study, consisting of speeches given in English at the European Parliament, and their interpreted versions in French. We have chosen to focus on one situational context, i.e. argumentative political discourse in the EU institutions, to avoid possible variations due to different contexts.

One sub-corpus consists of interventions at committee meetings, and another contains short speeches in the plenary. Since the rules of procedure place more stringent time constraints to speakers in the plenary, their succession is faster and the debates tend to be livelier. An additional corpus of two press conferences was used for comparison.

In order to study individual variation across speakers and interpreters, the corpus is organised in cross-tabulation groups: each speaker in a given group has been interpreted by every interpreter in the group, and vice versa, as summarised in Table 1.

Committee meetings			**Plenary sessions**		
Spk	**Int**	**Dur (s)**	**Spk**	**Int**	**Dur (s)**
S1en	I1fr	2 x 123	S4en	I3fr	2 x 330
S1en	I2fr	2 x 129	S4en	I4fr	2 x 228
S2en	I1fr	2 x 178	S5en	I3fr	2 x 170
S2en	I2fr	2 x 142	S5en	I4fr	2 x 126
S2en	I1fr	2 x 155	**Press conferences**		
S3en	I2fr	2 x 225	S6en	I5fr	2 x 205
S3en	I1fr	2 x 186	S6en	I6fr	2 x 273

Table 1: Corpus design and sample durations

Each of the 13 corpus samples is a time-synchronised recording of the original (EN) and the interpretation (FR). The total corpus duration is 82.5 minutes (2474 s or 41.3 min per language), with 6 different speakers (5 male, 1 female) and 6 different interpreters (2 male, 4 female).

2.2. Corpus annotation

All corpus samples have been transcribed, phonetised and aligned to the phone level using *SPPAS* (Bigi 2012) and *EasyAlign* (Goldman 2008). We performed syllabification with *SPPAS* for French, and our own reimplementation of the *P2TK syllabifier* for English. The phonetic alignment was manually corrected. The data is stored in *Praat* (Boersma & Weenink 2009) textgrids.

Prosodic information was extracted by applying f_0 stylisation using *ProsoGram* (Mertens 2004), *ProsoReport* (Goldman et al. 2011), and automatic prominent syllable detection (Goldman et al. 2012).

Bi-text alignment based on translational equivalence was performed on pause-separated units. This results in parallel macro-units of English speech and the corresponding French interpretation.

We have developed software to manage such a parallel spoken corpus, perform automated analyses and visualise the results, partly based on the open source project *Sonic Visualiser* (Cannam et al. 2010).

3. Temporal features of SI

The temporal organisation of the interpreter's speech depends upon the speaker and the reformulation process. The interpreter has to obtain enough source language input before starting to produce a coherent message in the target language. Gile (2009: 200) argues that interpreters use several strategies to cope with the cognitive load of this process, including stalling ('delaying the response'), varying the ear-voice span and their speech rate, and anticipating. These strategies are reflected in the observed temporal features of interpreters' speech.

Overall, interpreters make longer and less silent pauses than speakers. The distribution of silent pause durations is presented in Figure 1.

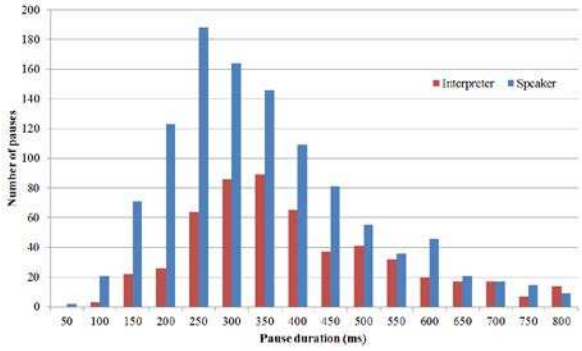

Figure 1: Silent pause duration distribution (blue: speaker / red: interpreter). All situations.

We also observe differences based on the situation: in the quicker debates at the plenary, the number of pauses and their average duration is smaller for speakers and interpreters alike, but the different patterns persist. At the press conferences, the speaker made more pauses than the interpreter, with similar pause durations.

	Mean (ms)	Medn (ms)	Std Dev	Avg num pauses /min
Speaker	**322**	**304**	**157**	**25.7**
Committee	328	311	152	20.3
Plenary	295	249	157	37.6
Press conf.	510	479	265	20.8
Interpreter	**594**	**379**	**640**	**13.8**
Committee	707	445	776	12.4
Plenary	412	329	319	18.2
Press conf.	827	508	962	10.0

Table 2: Silent pause length and frequency

Following the methodology in Goldman et al. (2010), the silent pause length distributions were modelled using a Gaussian mixture model in log time. Both were found to be bi-modal using a BIC criterion. The models are as follows (all times in ms):
For speakers (λ_1=8%, μ_1=105, σ_1=1.39 and λ_2=92%, μ_2=314, σ_2=1.49)

For interpreters (λ_1=45%, μ_1=329, σ_1=1.31 and λ_2=55%, μ_2=569, σ_2=2.26).

We also observed that in most (8 in 11) samples, and overall, the variance of speech rate was greater for speakers than for interpreters, which suggests that interpreters accelerate and decelerate more than speakers.

4. Global prosodic profile

	Committee				Plenary	
	S1	**S2**	**S3**		**S4**	**S5**
Agitation (semitones / second)						
Spk	7.60	8.35	6.60	**Spk**	4.80	7.20
I1	4.70	5.40	4.50	**I3**	4.40	4.90
Spk	7.10	8.80	4.40	**Spk**	4.90	7.30
I2	6.20	6.60	6.30	**I4**	5.80	5.60
F0 range (5%-95% interquartile range, semitones)						
Spk	7.80	12.30	12.10	**Spk**	11.10	12.90
I1	7.30	7.30	7.00	**I3**	8.20	9.40
Spk	7.40	15.00	7.90	**Spk**	11.20	13.70
I2	18.70	9.40	20.40	**I4**	5.70	6.20
Articulation ratio (%)						
Spk	84.7	89.3	81.9	**Spk**	79.9	79.4
I1	90.8	88.3	83.3	**I3**	82.8	81.8
Spk	84.1	89.2	81.7	**Spk**	79.6	79.0
I2	78.0	83.2	68.0	**I4**	90.5	92.6
Articulation rate (syllables / s, excluding pauses)						
Spk	6.10	4.95	4.90	**Spk**	5.10	5.40
I1	4.20	3.95	3.70	**I3**	4.80	4.80
Spk	5.70	5.10	4.60	**Spk**	5.70	5.00
I2	3.60	3.60	3.70	**I4**	4.90	4.40
Speech rate (syllables / s, including pauses)						
Spk	5.12	4.40	4.01	**Spk**	4.09	4.28
I1	3.83	3.48	3.06	**I3**	3.96	3.91
Spk	4.80	4.57	3.76	**Spk**	4.52	3.95
I2	2.78	2.98	2.51	**I4**	4.48	4.03

Table 3: Global prosodic measures

Interpreters have a lower average speech rate (including pauses) and articulation rate (excluding pauses) than speakers at the plenary and at committees, while these two rates are roughly equal at press conference. Interpreters' speech rate and articulation rate have greater variance than the corresponding rates for speakers. Most interpreters displayed a narrower f_0 range than speakers, and a lower melodic agitation (as defined in Goldman et al. 2007: 225), which suggests that they do not emphasise their speech as much as speakers do.

Finally, we can observe that within these general patterns, individual interpreters have their own speaking style. For example, I2 systematically uses a broad pitch range, regardless of the speaker. These observations lead us to study the dynamic evolution prosodic features, comparing the speaker and the interpreter over time.

5. Similarity and convergence

5.1 Methodology

We have studied the evolution of three prosodic features (speech rate, mean pitch and pitch range) over time based on the Time-Aligned Moving Average (TAMA) method, described by Kousidis et al. (2009). This method has been used to identify similarity and convergence in spontaneous dialogues. De Looze & Rauzy (2011) propose measures for synchrony (two speakers exhibiting similar speech patterns) and convergence (moving towards similar prosodic features over the course of time).

In simultaneous interpreting the two time series are naturally aligned, and since there is no interaction, only the interpreter can converge towards the prosodic features of the speaker.

For each prosodic feature under study, its value was extracted every 1 s, creating two time series that were normalised using the z-transformation. The moving average was calculated with a window size of 10 s. The synchrony (S) and convergence (C) strengths are calculated using Pearson's rho:

$$S = \rho_{\text{Pearson}}(x_1, x_2), \quad C = -\rho_{\text{Pearson}}(|x_1 - x_2|, t)$$

where x_1 and x_2 are the normalised TAMA values. Synchrony and convergence are independent, giving seven possible states: 3 states of similarity, 3 states of anti-similarity and 1 state of no similarity; see De Looze & Rauzy (2011). The dynamic nature of the phenomenon is captured by observing the plots of the two series and the two indices (S, C) over time.

5.2 Regions of synchrony

The interpreter and the speaker may be in synchrony on one prosodic parameter (e.g. speech rate) and at the same time in anti-synchrony on another one (e.g. pitch range). Table 4 summarises the percentage of time each state was observed.

	Speech rate			Pitch range		
	ANT	**NIL**	**SYN**	**ANT**	**NIL**	**SYN**
S1-I1	23	46	31	19	36	45
S2-I1	25	49	26	17	46	37
S3-I1	30	57	13	24	54	22
S1-I2	28	31	41	28	40	32
S2-I2	11	25	64	33	58	9
S3-I2	29	39	31	37	27	36
S4-I3	37	30	34	15	44	41
S5-I3	15	65	20	16	44	40
S4-I4	10	55	36	15	49	36
S5-I4	46	27	27	10	46	44

Table 4: Percentage of time (%) the speaker and the interpreter are in synchrony (SYN) or anti-synchrony (ANT) phases ($p<0.05$)

5.3. Speech rate

Speech rate presents the most interesting variation, as it is linked with the temporal constraints of simultaneous interpreting. We find alternating regions of synchrony (i.e. the interpreter is following the changes in speech rate of the speaker) and regions of anti-synchrony.

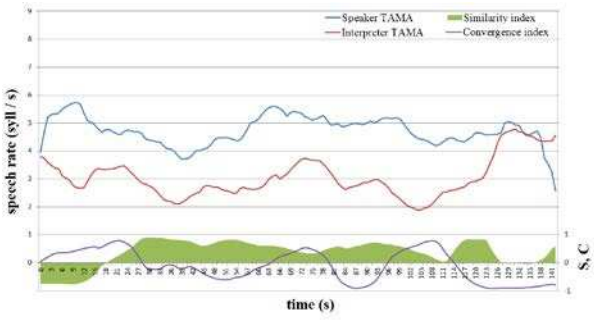

Figure 2: Speech rate TAMA plot with the interpreting following the speaker (S2/I2).

Two distinct styles emerge: the interpreter may be following quite faithfully the accelerations and decelerations of the speaker (e.g. Figure 2) or may be keeping their own pace (e.g. Figure 3). The second style was observed in most of the samples.

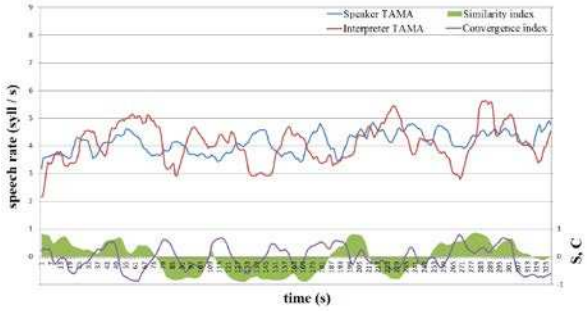

Figure 3: Speech rate TAMA plot with the interpreting 'chasing' the speaker but following their own pace (S4/I3)

5.4 Pitch range

In all samples, a high degree of synchrony in pitch range was observed (see Table 4). However, the interpreter's pitch range is smaller than the speaker's. This suggests that interpreters limit the extent to which they produce emphatic speech.

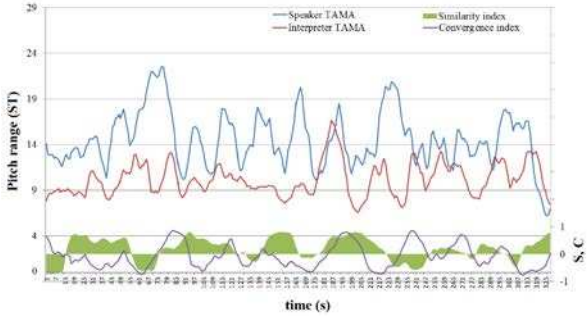

Figure 4: A typical pitch range TAMA plot

5.5. Mean pitch

Although the normalised time series are mostly similar (during 39% to 69% of the time), we do not observe convergence to the speaker's pitch, neither in absolute or relative terms. This seems logical since mean pitch imitation has been linked to mimicry in previous research (Couper-Kuhlen 1996).

6. Conclusions and further work

In conclusion, interpreters seem to adopt less expressive (more 'continuous') speaking style than the parliamentary speakers, a

manner that can be accounted for by the fact that their speech is dependent on, and subordinated to the speech of the orator. They seem to have their own strategies for responding to important changes in the speaker's prosodic cues, while at the same time exhibiting a distinct prosodic profile. The practical and cognitive constraints lead to a particular temporal organisation of their speech, with longer and less frequent silent pauses and a more variable speech rate.

We plan to expand this study, controlling for directionality (EN to FR, and FR to EN) and studying the relationship between the prosodic features and the structure of the original and the interpreters' speech.

Acknowledgments

I would like to express my gratitude to Prof. Anne-Catherine Simon for the guidance and encouragement she provided throughout this research.

References

Boersma, P. & Weenink, D. (2009). Praat: doing phonetics by computer. http://www.praat.org

Bigi, B. (2012). SPPAS: a tool for the phonetic segmentation of speech. *The 8th LREC*, Istanbul (Turkey), May 2012.

Cannam, C., Landone, C., Sandler M. (2010). Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files, *Proceedings of the ACM Multimedia 2010 International Conference*, pp. 1467-1468.

Couper-Kuhlen, E. (1996). The prosody of repetition: on quoting and mimicry. E. Couper-Kuhlen & M. Selting (eds.), *Prosody in conversation, studies in interactional sociolinguistics,* Cambridge University Press, pp. 366-405.

De Looze C., & Rauzy S. (2011). Measuring speakers' similarity in speech by means of prosodic cues: methods and potential. *Interspeech 2011*. pp. 1393-1396.

Eskenazi, M. (1993). Trends in speaking styles research, ISCA, pp. 501-509.

Gile, D. (2009). Basic concepts and models for interpreter and translator training (revised ed.) John Benjamins, Amsterdam/ Philadelphia.

Goldman, J.-Ph., Auchlin, A., Simon, A.C. & Avanzi, M. (2007). Phonostylographe: un outil de description prosodique. Comparaison du style radiophonique et lu. *Nouveaux cahiers de linguistique française* 28, pp. 219-237.

Goldman, J.-Ph. (2008). EasyAlign: a semi-automatic phonetic alignment tool under Praat, http://latlcui.unige.ch/phonetique

Goldman, J.-P., François, T., Roekhaut, S., Simon, A.-C. (2010): Étude statistique de la durée pausale dans différents styles de parole. Association Francophone de la Communication Parlée (ed.): *Actes des 28èmes Journées d'Étude sur la Parole. JEP 2010*. Mons, Belgium, 25-28 May 2010

Goldman, J.-Ph., Auchlin, A. & Simon, A.C. (2011). Description prosodique semi-automatique et discrimination des styles de parole. Yoo, H-Y & Delais-Roussarie, E. (eds.), *Actes d'IDP 2009*, Paris, Septembre 2009, ISSN 2114-7612, pp. 207-221.

Goldman, J.-Ph., Avanzi, M., Simon, A.C., Auchlin, A. (2012). A continuous prominence score based on acoustic features. *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, 9-13 September 2012.

Kousidis, S., Dorran, D., Mcdonnell, C. and Coyle,E. (2009). Convergence in human dialogues. Time Sries Analysis of Acoustic Features, *Proceedings of SPECOM 2009*, St. Petersburg, p. 2.

Léon, P. (1993). Précis de phonostylistique, Parole et expressivité, Nathan Université, Paris.

Llisteri, J. (1992). Speaking styles in speech research, *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, Dublin, Ireland, p. 28

Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. Bel, B. & Marlien, I. (eds.), *Proceedings of Speech Prosody 2004*, Nara (Japan), 23-26 March. ISBN 2-9518233-1-2.

Pöchhacker, F. (2004). Introducing interpreting studies. Routledge, London / New York.

Simon, A.-C., Avanzi, M., Goldman, J.-Ph. (2008). La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique, *Congrès Mondial de Linguistique Française 2008*, p. 151, DOI: 10.1051/cmlf08256

Simon A.C., A. Auchlin, M. Avanzi & Goldman, J.-Ph.(2010). Les phonostyles: une description prosodique des styles de parole en français. Abecassis, M. & G. Ledegen (eds.), *Les voix des Français. En parlant, enécrivant*, Peter Lang, Berne, pp. 71-88.