PAN ARCHIVES OF ACOUSTICS
Vol. **38**, No. 1, pp. 63–73 (2013)

DOI: 10.2478/aoa-2013-0008

Acoustic Features of Filled Pauses in Polish Task-Oriented Dialogues

Maciej KARPIŃSKI

Institute of Linguistics, Adam Mickiewicz University
al. Niepodległości 4, 61-874 Poznań, Poland; e-mail: maciej.karpinski@amu.edu.pl

(*received October 8, 2012; accepted January 15, 2013*)

Filled pauses (FPs) have proved to be more than valuable cues to speech production processes and important units in discourse analysis. Some aspects of their form and occurrence patterns have been shown to be speaker- and language-specific. In the present study, basic acoustic properties of FPs in Polish task-oriented dialogues are explored. A set of FPs was extracted from a corpus of twenty task-oriented dialogues on the basis of available annotations. After initial scrutiny and selection, a subset of the signals underwent a series of pitch, formant frequency and voice quality analyses. A significant amount of variation found in the realisations of FPs justifies their potential application in speaker recognition systems. Regular monosegmental FPs were confirmed to show relatively stable basic acoustic parameters, which allows for their easy identification and measurements but it may result in less significant differences among the speakers.

Keywords: filled pauses, paralinguistics, dialogue, acoustic properties, Polish.

1. Filled pauses as paralinguistic components of spoken utterances

Paralinguistic features of utterances (henceforth PLFs) were extensively explored by linguists early in the 1960s (Trager, 1960, 1961, 1964; Crystal, 1963, 1966, 1974, 1975). Since then, they have been acknowledged as an indispensible component of spoken communication. They may reveal important facts on the speaker him/herself, including his/her age, gender, origin, social background or education, as well as his/her present emotional state, attitude towards the topic of conversation or towards the conversational partner (Gobl, Ni Chasaide, 2003; Wallbott, Scherer, 1979; Pakosz, 1982; Bortfeld, 2001). Some of them may also provide cues in the analysis of the process of speech production (e.g., disfluencies (Fromkin, 1971, 1973)). PLFs often contribute to an individual, idiosyncratic speaking style but most people are also able to control many of them consciously.

As PLFs form an extraordinarily heterogeneous group, it is difficult to cover all of them with a single definition. Moreover, they can be defined only as precisely as precise the boundaries of language can be. As a result, researchers tend to define them by enumeration and the inventories proposed in literature are often selective. Among the most frequently mentioned examples of PLFs, there are prosodic features and voice quality parameters (Crystal, 1966). Some of them can be relatively easily measured as local acoustic parameters of speech signal (energy, pitch frequency). Others can be determined only when longer stretches of speech are analysed. Finally, some of them have more "structural" nature. For example, in an emotional utterance, words may be sequenced in an atypical way, accentuation may be also uncommon. But this can be only noticed once the entire utterance is taken into account and compared to some other typical or common structures.

In recent decades, many studies of paralinguistic features have been focused on potential cues to the emotional or attitudinal value of utterances in psychological and psycholinguistic (e.g., Ladd, 1985; Patel *et al.*, 2011; Johnstone, Scherer, 2000), communicational (e.g., Benuš, 2009; Gravano, 2011) or technological (Campbell, 2002; ten Bosh, 2003) contexts. Social and intercultural studies of paralinguistic features have also brought interesting results (Abelin, Alwood, 2000; Doddington, 2001; Abelin, 2008; Burkhardt *et al.*, 2006; Scherer *et al.*, 2000; Campbell, 2004, 2007), especially in relation to emotionality. But in everyday communicative situations, one rarely faces vividly emotional speech. Subtle background emotions and the mixtures of emotions driv-

ing human everyday behaviour may be are extremely difficult to detect and decode. Another area of research related to the form and function of PLFs has been speaker identification and recognition. The variation of the acoustic form of utterances poses additional problems in the area of speech recognition, but it makes automatic speaker recognition and identification possible (Mary, Yegnanarayana, 2008; Schötz, 2002), especially – but not solely – in forensic applications (Gonzales-Rodriguez, 2008; Sacks, Koehler, 2005). The importance of paralinguistic features of utterances in language communication fully justifies attempts towards the formulation of spoken language grammar that would encompass paralinguistic information (Campbell, 2002) as well as attempts towards a comprehensive theory of disfluencies (Shriberg, 1994).

Silent and filled pauses (SPs and FPs, respectively) seem to be especially frequently explored categories of paralinguistic phenomena. It was found very early in the studies on spontaneous speech that FPs may signal problems in the process of lexical access (Maclay, Osgood, 1959; Goldman-Eisler, 1968) as well as in syntactic processing (Boomer, 1965). More recent views on the origin and role of FPs are presented, for example, by Shriberg (2001) or Ward (2005). Their distribution, length and form say much about the pragmatic aspects of utterances as well as about the speaker him/herself. There is no doubt that pauses have certain communicative value (e.g., Saville-Troike, 1985; Local, Kelly, 1986; Nishinuma, Hayashi, 2004). They may function as markers of discourse structure (Swerts, 1998) and can be consciously used as a stylistic device but they may also reveal a peculiar, idiosyncratic speaking style. As opposite to monologue speech, in dialogues, their occurrence may also result from a wide variety of interaction-related factors, e.g. processing input from the conversational partner, formulating a reply, waiting for an appropriate moment for turn transition, as well as from other aspects of alignment tendencies.

FPs may be realised in a variety of ways – not only as centralised vowels or "creaky sounds" but also as full lexical units (sometimes taboo words) that are, however, used for purposes different from exploiting their lexical meaning. In such cases, they may be marked by a peculiar pronounciation, involving atypical lengthenings or pitch changes.

In their study of eight languages, Candea *et al.* (2005) found that non-lexical, vocalic fillers were not language-specific in terms of duration or pitch frequency but rather in terms of vocalic quality and segmental structure. Similar findings were reported by Vasilescu *et al.* (2004, 2005) who pointed to some language specific differences in the values of f_1 and f_2. Stepanova (2007) found significant inter-speaker variation in the realisations of FPs in Russian. Duez (2001) reports the f_0 values of filled-pause onsets to be stable within the same speaker as they are "linked to the absolute, physiological aspects of speech".

These and some other studies show that (1) FPs are relatively easy to detect in the stream of speech on the basis of their acoustic properties; (2) FPs show certain speaker-specific features, both in terms of occurrence patterns and acoustic realisations. As a consequence, FPs are potentially useful in automatic speaker identification or recognition.

Early studies of FPs and SPs in semi-spontaneous Polish utterances (Francuzik *et al.*, 2002; Karpiński, 2006, 2007) not only show differences in the distributions of their durations (which is intuitively obvious as SPs – unlike FPs – can be arbitrarily long) but also their co-occurrence patterns. While the form of their realisation has been briefly discussed in (Karpiński, 2008), their acoustic properties have never been studied in depth for the Polish language.

2. The aims of the study

In the present study, selected acoustic properties of FPs in Polish are explored in order to find the areas of individual differences that can be potentially useful in speaker identification and recognition or just as markers of individual speaking style.

Speaker recognition and identification are intensively developing areas of research due to their potential applications in security systems, access management systems and forensics (Beigi, 2011a, 2011b). Each human has an uniquely shaped vocal tract which contributes to the individual acoustic features of voice. Still, it is not easy to define the "vocal print" (voice print) which would be a set of acoustic features that allow for unambiguous identification of the speaker. Most researchers seem to follow the path of gathering possibly numerous measuremets of parameters that can be extracted from the acoustic signal and then, with the use of advanced statistical methods, looking for their most efficient combinations and hierarchies that can be used in the procedure of speaker identification or recognition (Beigi, 2011b). However, as the efficiency of such an approach can be significantly limited by the quality of available voice recordings, another path is to include idiosyncratic structural-linguistic properties of utterances – e.g., the frequency of words or phrases, or the way of building sentences (e.g., Doddington, 2001; Shriberg, 2007) but also any potentially speaker-specific paralinguistic features and phenomena, including SPs and FPs.

The analyses are based on dialogue recordings as conversational speech shows peculiar properties and its paralinguistic profile may significantly differ from what is found in monologues. In general, one may expect more disfluences (including FPs) than in monologues as the cognitive load related to the interactivity

normally seems to be higher than in the case of monologue where the speaker may prepare some portions of her/his talk beforehand and does not face interruptions nor fighting for the floor. Here, instrumental analyses are focused on the pitch frequency, first two formant frequencies as well as on jitter and shimmer measures in simple monosegmental fillers. Detailed occurrence patterns of FPs are not covered by the present study because the data in hand are still to sparse and the number of factors related to interactivity and dialogue-specific environment is to high to build a comprehensive model.

3. Material under study: DiaGest2 Corpus

DiaGest2 multimodal corpus (e.g., Karpiński, Jarmołowicz-Nowikow, 2010) consists of twenty task-oriented "origami" dialogue session recordings, each of approximately five minute duration. The task of the participants was to reconstruct a figure made of paper. The figure was fully visible to the instruction giver (IG) and not visible to the instruction follower (IF) who was provided with all the necessary materials for its reconstruction. In ten sessions, IF and IG faced and could see each other (mutual visibility condition, MVC), while in another ten sessions they could not see each other (limited visibility condition, LVC). IGs were gender balanced (both in the MVC and LVC), while most of the IFs were females. Each participant took part in the task only in one of the conditions and only in one role (IF or IG).

FPs were manually tagged in the process of transcription and segmentation using Praat (Boersma, Weenink, 2012), both in the word and syllable tiers, and automatically extracted from dialogue recordings. As the number of utterances and FPs produced by IFs was much lower, only the utterances by IGs are analysed here. IFs produced less utterances and most of them were much shorter and built of shorter phrases than in the case of IGs so there were less filler-evoking contexts. The total numbers of IG's fillers of FPs in the MVC and LVC sessions was almost equal (Table 1), although there were clear individual differences among speakers. The total number of FPs produced by IGs ranged from 7 to 61.

Table 1. The frequencies of FPs in the DiaGest2 corpus.

Signals	MVC	LVC	Sum
Extracted	255	256	511
Female	143	142	285
Male	112	114	226
Excluded from further analyses	82	41	123

A significant proportion of FPs had to be excluded from further instrumental analyses after first audition. Many of them included noises or overlapped speech, and some were realised as voiceless, breathy sounds. Those realised in creaky voice (ca. 3%) and as nasal(ised) sounds (ca. 5.3%) were also rejected. Finally, some of them were compound sequences of sounds which did not match the analytical framework adopted for this study. As a result, the initial set of signals was significantly reduced but it still allowed for a number of measurements and analyses. This limitation has also its advantages: "Clean" FPs can be more easily traced and extracted automatically so the set under study is more compatible with the potential input data for FPs-based speaker recognition computer systems.

Although only the data from IGs will be analysed in this study, in Fig. 1, the "FPs per syllable" rate is shown for both IGs and IFs in the MVC (calculated by dividing the number of FPs by the number of syllables). On average, the rates are higher for IGs which can be easily justified by a higher complexity of this role. However, the data show that there can be strong individual variation in the FP per syllable rate (e.g., see Session 9 and 10). In any case, the proportions reflect the fact that the cognitive load of this particular task was quite different for IGs and IFs. IGs produced more complex phrases, forming longer sequences while IFs could focus on understanding, guessing and giving feedback.

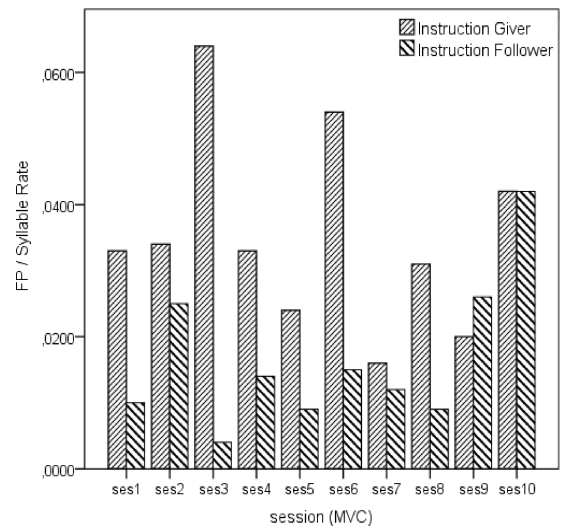

Fig. 1. "Filled pauses per syllable" rates for IGs and IFs in the MV condition.

4. Acoustic properties of FPs

Acoustically homogenous FPs are relatively easy to track down automatically in the stream of speech which makes them convenient for automatic sampling and analysis. Although they seem to be relatively stable in terms of pitch frequency and voice quality, they may still have enough of individual variance for speaker identification or recognition. Speakers use idiosyncratic "lexical" fillers so they may also tend to shape their

non-lexical fillers volitionally in a peculiar way. Obviously, the acoustic characteristics of fillers as mostly vocalic sounds must be strongly influenced by the unique shape of each vocal tract.

4.1. Durations of FPs

The durations of fillers vary extremely. The distribution presented in Fig. 2 shows covers only the analysed subset of "clean" signals. It may be slightly different from a distribution of the entire set of FPs as it does not include compound fillers – not very numerous but, as a rule, being much longer. The distribution of FPs duration was similar to that obtained by Francuzik *et al.* (2002). It was also skewed to the left but here the maximum is around 200–300 ms while in the cited work it reached ca. 500 ms.

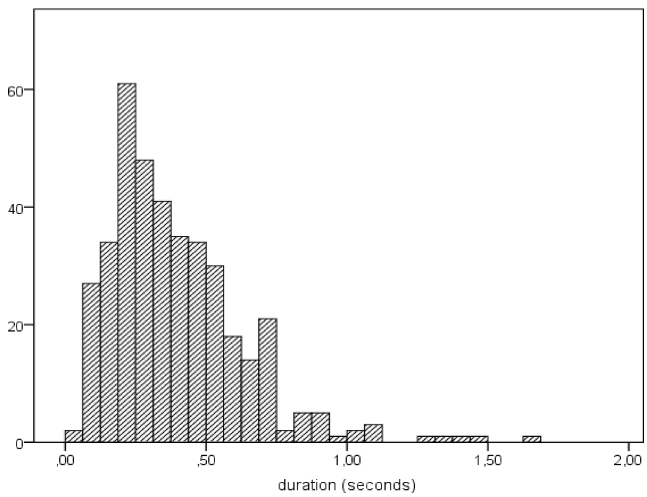

Fig. 2. The distribution of FP's duration in the utterances produced by IGs.

The difference between the mean values of FP duration in female and male speakers turned out to be statistically insignificant ($p > 0.05$). However, one-way Anova performed on log-transformed data (to compensate for the skew) showed significant differences between the mean durations of FPs for individual speakers ($F = 3.43$, $p < 0.01$; $df = 18$ as one of the male voices has a very limited number of observations and was excluded from this and further Anova calculations).

4.2. Pitch frequency in FPs

The values of the mean pitch frequency averaged for the entire gender groups and visibility conditions are presented in Table 1. For female speakers, pitch frequency of FPs was significantly higher in the MVC (t-test, $p = 0.001$) and there was no such a difference difference for males. While the number of subjects does not allow for brave hypothesising, one may understand this result as not necessarily intuitive. One of possible explanations may be that the mutual visibility condition evokes more emotional speech in females (but not in males) and leads to higher pitch values. In Table 2, pitch range data are presented for both female and male speakers in both the conditions.

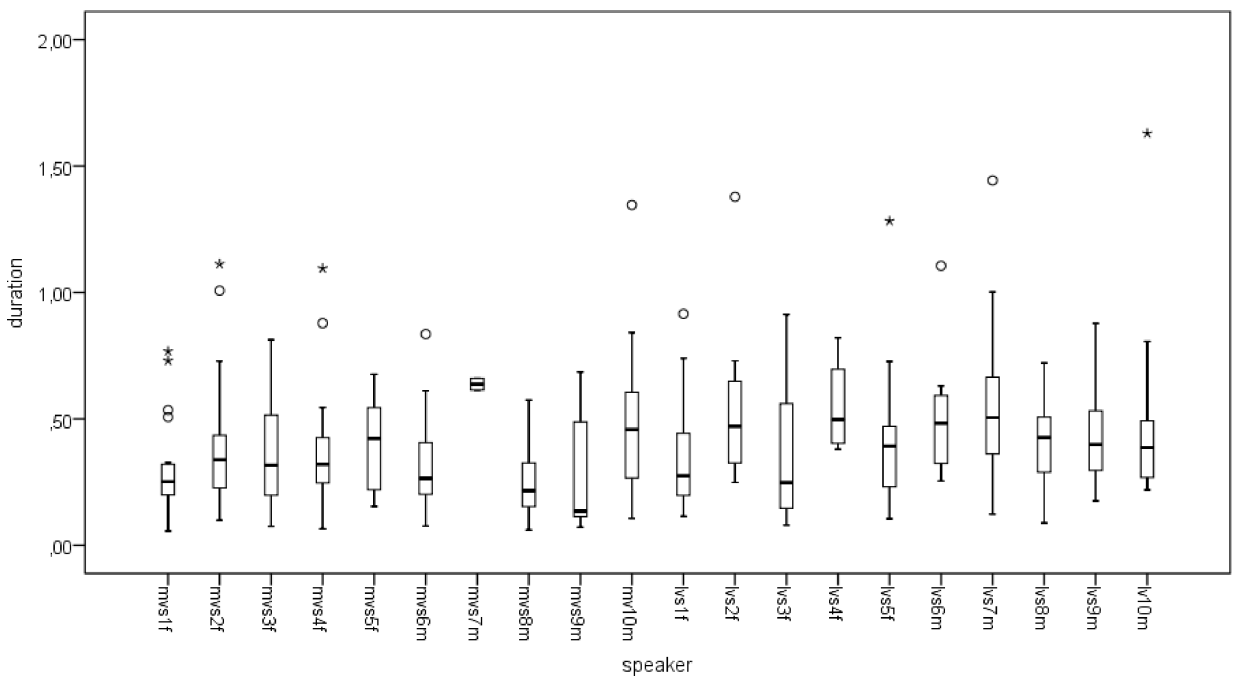

Fig. 3. Individual variation in FP duration. Speakers grouped according to the recording condition (mv – mutual visibility condition, lv – limited visibility condition, f – female, m – male).

Table 2. Mean pitch frequency in female and male voices in both the conditions (minimum and maximum values calculated directly from individual signals).

condition	MVC		LVC	
gender	female	male	female	male
average pitch frequency [Hz]	229	124	219	141
standard deviation (pitch freq.)	22	17	26	22
maximum pitch frequency [Hz]	278	193	120	66
minimum pitch frequency [Hz]	169	84	317	198

Table 3. Relative pitch frequency range for male and female speakers in the two conditions (MVC and LVC) calculated as $(f_{0\,\max} - f_{0\,\min})/f_{0\,\max}$.

condition	MV		LV	
gender	female	male	female	male
mean pitch frequency change	0.04	0.02	0.07	0.04
standard deviation	0.07	0.04	0.08	0.07
maximum pitch frequency change	0.30	0.28	0.42	0.43

As it can be seen in Fig. 5, the mean relative pitch change is also a parameter which may take very individual values (one-way Anova, IGs in two conditions, $p < 0.001$). On the other hand, the majority of mean values are in a limited range. There are more "wide range" speakers among females and there is a signi-

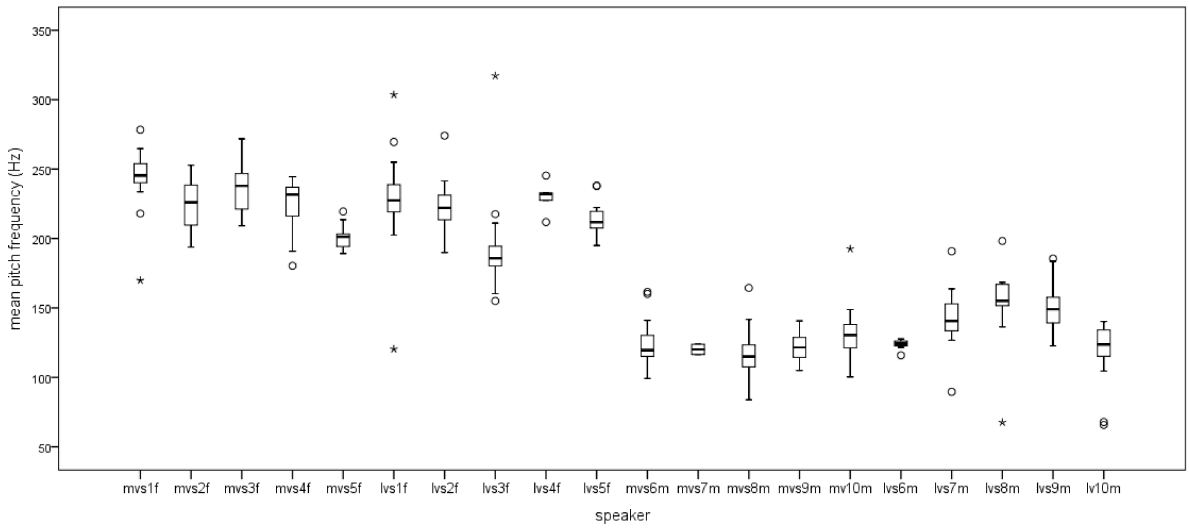

Fig. 4. Mean pitch frequency in FPs realised by twenty speakers (speakers grouped by gender; mv and lv stand for MVC and LVC, respectively).

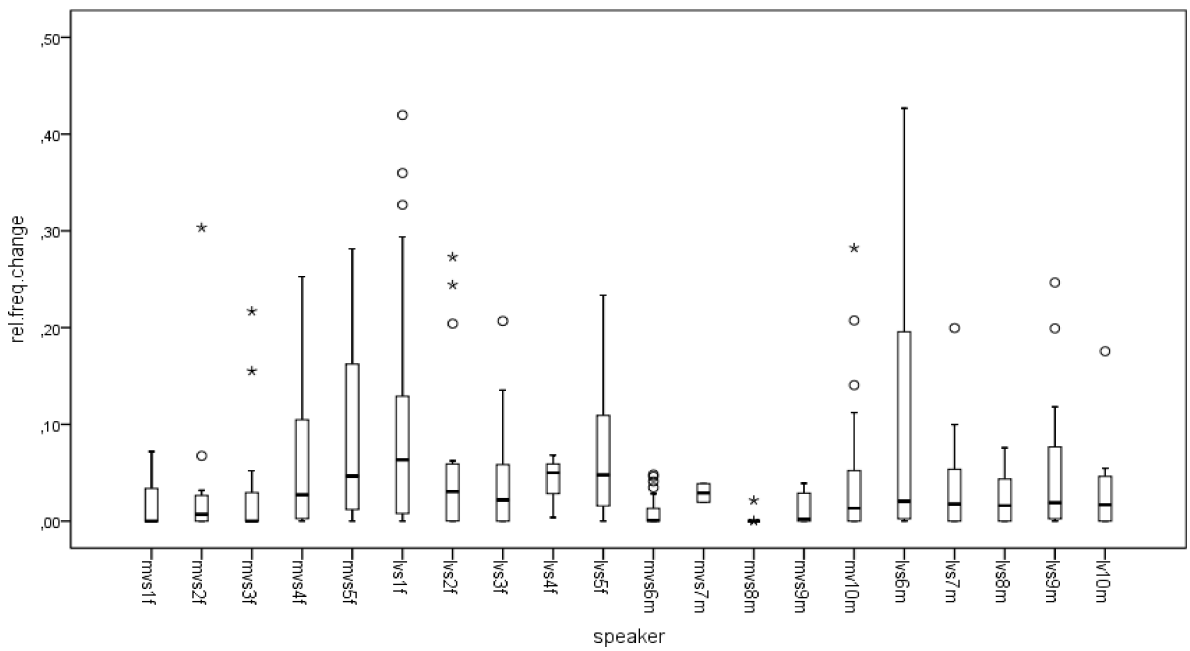

Fig. 5. Mean pitch frequency range within FPs for all the speakers (grouped by gender).

ficant difference between the means for females and males ($p < 0.001$). However, there is no significant difference between the values both for female and male speakers in the two conditions ($p > 0.05$).

4.3. Voice quality in FPs

Voice quality is often understood as a subjective perceptual parameter and, as such, it is difficult to express in terms of acoustic measures of the signal itself. More precisely, it was defined as the auditory coloring of voice, derived from a variety of laryngeal and supralaryngeal features (Abercrombie, 1967, p. 91; Laver, 1980, p. 1; Trask, 1996, p. 381; Keller, 2005). Still, it remains a confusing term as it encompasses phenomena of various origin and character. Voice quality in FPs has been found to be language specific (Shriberg, 2001, p. 156). It has been also demonstrated to have some functions in dialogue interaction (Local, Kelly, 1986; Ogden, 2001). Among the most frequently used acoustic parameters related to voice quality are jitter (corresponding to the variation in the pitch frequency) and shimmer (corresponding to the variation of amplitude). Both are measures of phonation irregularities and their high values may signal pathology. In a number of studies, the usefullness of jitter and shimmer measurements in speaker recognition or identification has been indicated (Farrús, Hernando, Ejarque, 2007; Farrús, Hernando, 2009). They have been shown to be related to the speaker's age, gender, smoking habits, and some other features (Ludlow *et al.*, 1982).

One can assume that voice quality parameters may change during an utterance, depending on the phonetic fundament and context (segmental, suprasegmental) and as well as possible emotional factors or even changes in the articulatory effort in longer stretches of speech. Still, in the case of monosegmental nonlexical fillers one may expect them to be relatively stable (Audhkhasi *et al.*, 2009). Therefore, they may be especially precious as a source of data on basic, presumably speaker-dependent voice quality parameters.

In the present study five jitter (local, local absolute, rap, ppq5, DDP) and six shimmer measures (local, local dB, DDA apq3, apq5, apq11) were extracted using Praat. Their basic descriptions can be found, e.g., in (Farrús *et al.*, 2007) or (Rusz *et al.*, 2011), as well as in Praat Manual. Although the measurements were carried out on a reduced set of preselected, "clean" signals, the number of analysis errors was significant, especially for apq5 and apq11 shimmer. As a consequence, only some of the measured variables were taken into account in further steps and, in some cases, two of the speakers were excluded as having too few FPs that could be analysed.

Among jitter measures, only the mean values of local and local absolute jitter were found significantly different for female and male speakers ($p < 0.05$; $df = 18$ as one of the voices was excluded). For shimmer, only the mean values of shimmer DDA were significantly different for the two genders ($p < 0.05$). One-way Anova showed significant differences in the mean values for individual speakers in the case of local absolute jitter ($p < 0.01$) as well as for local and PPQ5 jitter ($p < 0.05$). Similar Anova tests were conducted for four shimmer measures but the differences among the means for respective subjects were signicant only for DDA and APP3 ($p < 0.05$).

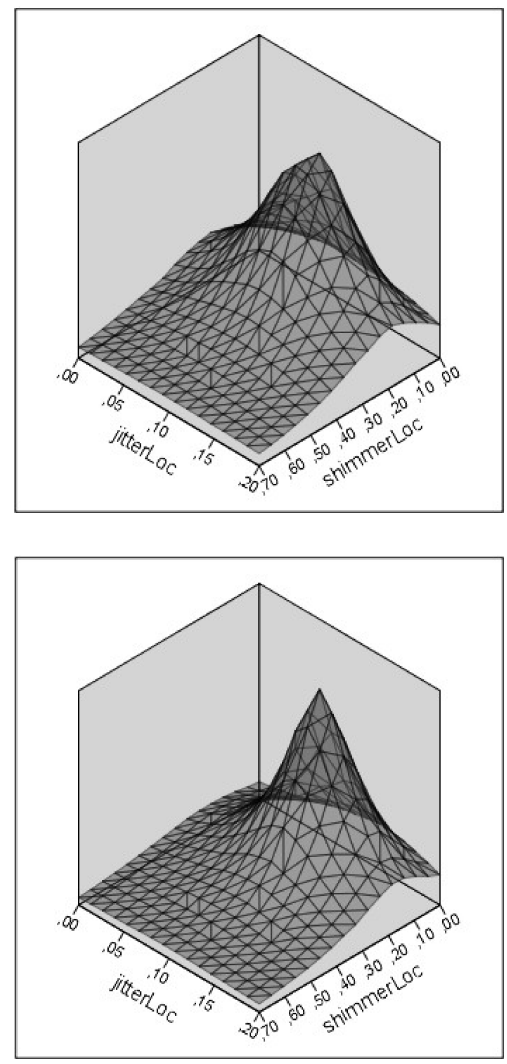

Fig. 6. The distribution of local jitter and shimmer values for female (top panel) and male (bottom panel) speakers.

4.4. Format frequencies in FPs

Formant frequencies are major cues to the identification of vocalic segments (Rosner, Pickering, 1994). For example, Albalá *et al.* (2009) show the variation of formants in vowels in Spanish that can be employed for the purpose of speaker identification. As most of the FPs under study are vocalic, their

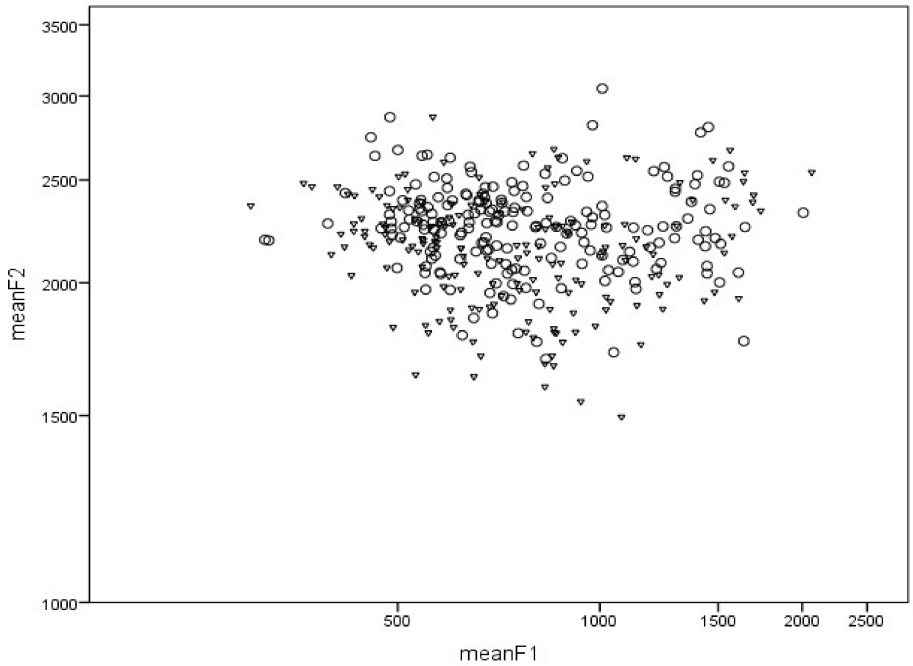

Fig. 7. Formants (f_1 and f_2) in FPs produced by female (circles) and male (triangles) speakers.

formant frequencies may bear some important information on their form. Vasilescu *et al.* (2005) argue for potentially language-specific character of f_1/f_2 formant values in FPs. Stepanova (2007) finds that FPs in Russian differ in terms of f_1 and f_2 values from vowels in stressed positions. She also points to speaker-specific characteristics of FPs and argues that it makes them useful in speaker recognition. Many methods used for the idenfication of FPs in the stream of speech are based on the measurements of formant values (Audhkhasi *et al.* 2009; Wu, Yan, 2004).

The material under study was analysed for the first two formant frequencies using the robust formant extraction algorithm in Praat. The signals that gave extreme or clearly erroneous results were excluded from further statistical analyses. Although the mean values of f_1 and f_2 were not significantly different in female and male speakers, Anova test showed significant differences in the mean formant values for individual voices ($df = 18$, $p < 0.01$).

Typical Polish monosegmental vocalic fillers are articulatorily less centralised than their English counterparts. Their formant values tend to group in the upper right area of the vocalic loop (cf. Jassem, 1973) but they can reach much further in terms of their f_1 values.

Table 4. Mean formant values for female and male speakers.

gender	female		male	
formant freq.	f_1	f_2	f_1	f_2
mean value [Hz]	875	2265	827	2156
st. deviation	382	220	362	257

4.5. Other findings

Some types of FPs were excluded from the above study as they could be hardly analysed for at least some of their acoustic parameters or they were realised as compound units that could not be covered by the present analytic framework. Among them, the following were most frequent:

- *Creaky fillers.* More typical of some speakers in the group but can be found in almost any voice. As in other cases, distorted phonation is more frequent in the final, low-energy stages of fillers but some FPs are realised solely in creaky voice.
- *Voiceless fillers* (sigh-like sounds). Rare but important as it seems that they may have different meaning from their fully voiced counterparts.
- *Compound fillers.* Built of at least two significantly different segments with a perceptually evident transition between the two (as in the case of, e.g., closing lips in the middle of the filler and producing its remaining part as a nasal sound).

These types of FPs would require a modified approach to acoustic measurements, involving different measures. There is no doubt that they deserve further studies. However, the assummption that the proportion of peculiar (creaky, voiceless, compound, etc.) FPs is typical of a given individual and may serve as a basis for speaker recognition or identification remains risky. It may prove dependant on the particular type of communicative situation or the mode of interaction. In any case, a substantially larger corpus, both in terms of the number of speakers and signals under analysis, would be necessary to explore these issues.

5. Conclusion and directions for further research

The data analysed in the present study come from a single, peculiar type of communicative situation. As a consequence, they are more coherent, which allows for the control of more variables and facilites statistical exploration even with a relatively limited number of speakers (from 18 to 20) and signals (ca. 400 used in most analyses). The initial rejection of distorted and imperfectly recorded signals may potentially introduce a bias as it might have reduced the number of samples coming from the most emotional exchanges, with many overlaps and from overloaded recordings or from the stages where the IF was most active and caused most noise manipulating the figure. However, it may actually reflect what can be gathered in real life circumstances where speech samples are recorded for the purpose of speaker identification and where similar quality issues may arise.

The acoustic characteristics of FPs in Polish seem to comply with many of the findings and claims cited above as well with those from some other studies (e.g., Stepanova, 2007; Wu, Yan, 2004; Vasilescu, Adda-Decker, 2007) but a more direct cross-linguistic comparison requires additional work. Polish FPs form a group showing homogeneity in a number of dimensions. While ca. 10% cases were excluded from instrumental analyses (nasalised, creaky, voiceless and compound fillers), the remaining 90% were relatively stable in terms of the analysed parameters, often had flat pitch contours and show little variation in the values of the first two formants.

The range of FPs duration was found to be extremely wide and bear some signs of speaker-dependance, although it was not gender-specific. The mean pitch frequency of the FPs in the mutual visibility condition (MVC) was found significantly higher than in the limited visibility condition (LVC) for female but not for male speakers. The mean values of the pitch range turned out to be significantly different in the set of twenty speakers but it may be due to some extreme individual means, with the remaining group kept within a rather limited range. Accordingly, pitch range alone would be not a clearly speaker-specific measure except for some cases. Audhkhasi *et al.* (2009) reports that their formant-based technique for filler detection is more accurate than cepstrum and pitch-based methods. Simultaneously, f_1 and f_2 values in the Polish FPs significantly differed in individual speakers. Accordingly, formant frequencies measurements may support both the identification of fillers in the stream of speech and can be expected to support speaker identification. Jitter-related voice parameters showed a limited level of distinctiveness – only the mean values of the local absolute, local and PPQ jitter as well as the DDA and APP3 shimmer turned out to be significantly different for the voices under study.

A significant proportion of the initial set of signals had to be rejected and new methods of exploration would be certainly needed for a systematic analysis of their acoustic properties different from pitch or formants but referring to some other spectral parameters. Another issue to be addressed in future research is the question of contextual influences on the acoustic realisations of FPs. Shriberg (1999) suggests that the pitch pattern of the filler is influenced by the intonational contour of the preceeding utterance but this can be only one among many other context-driven phenomena one should take into account. Finally, the structural properties of utterances should be taken into acount (e.g., the placement of a given FP in an utterance) as well as the correlation of speech production problems with certain gestural behaviour (Christenfeld, Schachter, Bilous, 1991; Esposito *et al.*, 2001; Jarmołowicz-Nowikow, Karpiński, 2012). For future studies, the data will be extended with a new corpus of task-oriented dialogues which will allow for a more reliable verification of hypotheses. Various configurations of the features will be explored in order to find optimum sets for the purpose of speaker recognition (cf. Demenko, 2000). Alternative, more flexible exploration techniques based on data mining will be also employed.

Acknowledgment

This work is supported from the financial resources for science in the years 2010–2012 by the National Centre for Research and Development as a development project O R00 0170 12.

References

1. Abelin Å., Allwood J. (2000), *Cross-linguistic Interpretation of Emotional Prosody*, ITR Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK.

2. Abelin Å. (2008), *Anger or Fear? – Crosscultural multimodal interpretations of emotional expressions*, [in:] *Emotions in the human voice*, Izdebski K. [Ed.], pp. 65–73, Vol. 1, Plural Publ. Co., San Diego.

3. Abercrombie D. (1967), *Elements of general phonetics*, Edinburgh University Press, Edinburgh.

4. Albalá M. J., Battaner E., Gil J., Llisterri J., Machuca M., Marrero V., Ríos A. (2009), *Vowel formant structure and speaker identification. A perceptual study*, CIP 2009– 3a Conferência Ibérica de Percepção, Guimarães, Portugal, 8-10 Julho 2009.

5. Audhkhasi K., Kandhway K., Deshmukh O., Verma A. (2009), *Formant-based technique for automatic filled-pause detection in spontaneous spoken English*, Proc. ICASSP, pp. 4857–4860, Taipei, Taiwan.

6. Beigi H. (2011a), *Speaker Recognition* [in:] *Biometrics*, Yang J. [Ed.], InTech. ISBN: 978-953-307-618-8, Available from: http://www.intechopen.com/books/biometrics/speaker-recognition.

7. Beigi H. (2011b), *Fundamentals of Speaker Recognition*, Springer, New York.

8. Benuš Š. (2009), *Variability and stability in collaborative dialogues: Turn-taking and filled pauses*, Proceedings of Interspeech 2009, pp. 709–799, Brighton.

9. Boersma P., Weenink D. (2013), *Praat: doing phonetics by computer* [Computer program], Version 5.3 (retrieved http://www.praat.org/).

10. Boomer D. S. (1965), *Hesitation and grammatical encoding*, Language and Speech, **8**, 148–158.

11. Bortfeld H., Leon S., Bloom J., Schober M., Brennan S. (2001), *Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender*, Language and Speech, **44**, 2, 123–147.

12. ten Bosch L. (2003), *Emotions, speech and the ASR framework*, Speech Communication, **40**, 1-2, 213–225.

13. Burkhardt F., Audibert N., Malatesta L., Türk O., Arslan L., Aubergé V. (2006), *Emotional Prosody – Does Culture Makes A Difference?*, Proceedings of Speech Prosody 2006 Conference, Dresden, Germany.

14. Campbell N. (2002), *Towards a grammar of spoken language: Incorporating paralinguistic information*, Proceedings ICSLP 2002, Denver, Colorado.

15. Campbell N. (2004), *Listening between the lines. A study of paralinguistic information carried by tone of voice*, pp. 13–16, International Symposium on Total Aspects of Languages TAL2004, Beijing, China.

16. Campbell N. (2007), *Whom we laugh affects how we laugh*, Proc. Workshop on "The Phonetics of Laughter", pp. 61–65, Saarbrücken, Germany.

17. Candea M., Vasilescu I., Adda-Decker M. (2005), *Inter- and intra-language acoustic analysis of autonomous fillers*, Proceedings of DISS05, Aix-en-Provence, France.

18. Christenfeld N., Schachter S., Bilous F. (1991), *Filled pauses and gestures: It's not coincidence*, Journal of Psycholinguistic Research, **20**, 1–10.

19. Crystal D. (1963), *A perspective for paralanguage*, Le Maître Phonétique, **120**, 25-29.

20. Crystal D. (1966), *The linguistic status of prosodic and paralinguistic features*, Proceedings of the University of Newcastle-upon Tyne Philosophical Society, **1**, 8, 93–108.

21. Crystal D. (1974), *Paralinguistics*, [in:] *Current trends in linguistics*, T. A. Sebeok [Ed.], **12**, pp. 265–295, Mouton, The Hague.

22. Crystal D. (1975), *Paralinguistics*, [in:] *The body as a medium of expression*, Benthall J., Polhemus T. (Eds.], pp. 162–174, Institute of Conteporary Arts, London.

23. Demenko G. (2000), *Analysis for suprasegmental features for speaker verification*, 8-th Australian International Conference on Speech Science and Technology, pp. 294–299, Canberra.

24. Doddington G. (2001), *Speaker recognition based on idiolectal differences between speakers*, Proceedings of the Eurospeech, **4**, 2521–2524.

25. Duez D. (2001), *Acoustic-phonetic Characteristics of Filled Pauses in Spontaneous French*, ITRW on Disfluency in Spontaneous Speech, pp. 41–44, Edinburgh.

26. Esposito A., McCullough K. E., Quek F. (2001), *Disfluencies in gesture: Gestural correlates to filled and unfilled speech pauses*, IEEE Workshop on Cues in Communication, Kauai, Hawaii.

27. Farrús M., Hernando J. (2009), *Using jitter and shimmer in speaker verification*, IET Signal Processing, **3**, 4, 247–257.

28. Farrús M., Hernando J., Ejarque P. (2007), *Jitter and shimmer measurements for speaker recognition*, pp. 778–781, Proceedings of Interspeech 2007 Conference, Antwerp, Belgium.

29. Francuzik K., Karpiński M., Kleśta J. (2002), *A Preliminary Study of the Intonational Phrase, Nuclear Melody and Pauses in Polish Semi-Spontaneous Narration*, Proceedings of Speech Prosody 2002 Conference, Aix-en-Provence, France.

30. Fromkin V. A. (1971), *The nonanomalous nature of anomalous utterances*, Language, **47**, 27–52.

31. Fromkin V. A. [Ed.] (1973), *Speech errors as linguistic evidence*, Mouton Publishers, The Hague.

32. Garg G., Ward N. (2006), *Detecting Filled Pauses in Tutorial Dialogs*, Technical Report UTEP-CS-06-32.

33. Gobl C., Ni Chasaide A. (2003), *The role of voice quality in communicating emotion, mood and attitude*, Speech Communication, **40**, 1-2, 189–212.

34. Goldman-Eisler F. (1968), *Psycholinguistics. Experiments in spontaneous speech*, The Academic Press, London and New York.

35. Gonzalez-Rodriguez J. (2008), *Forensic Automatic Speaker Recognition: Fiction or Science?*, Proceedings of Interspeech 2008, Brisbane, Australia.

36. Gravano A., Levitan R., Willson L., Benuš Š., Hirschberg J., Nenkova A. (2011), *Acoustic and Prosodic Correlates of Social Behavior*, Proceedings of Interspeech 2011, Florence, Italy.

37. Jarmołowicz-Nowikow E., Karpiński M. (2012), *The form and function of pointing gestures in task-oriented dialogues*, Conference of the International Society for Gesture Studies (Book of Abstracts), Lund, Sweden.

38. Jassem W. (1973), *Principles of Acoustic Phonetics*, [in Polish: *Podstawy fonetyki akustycznej*], PWN, Warszawa.

39. Johnstone T., Scherer K. R. (2000), *Vocal communication of emotion*, [in:] *The Handbook of Emotions*, 2nd Ed., Lewis M., Haviland J. [Eds.], pp. 226–235, Guilford, New York.

40. Karpiński M. (2006), *Structure and intonation of Polish task-oriented dialogue* [in Polish], Wydawnictwo Naukowe UAM, Poznań.

41. Karpiński M. (2007), *Selected quasi-lexical and non-lexical units in Polish map task dialogues*, Archives of Acoustics, **32**, 1, 51–65.

42. Karpiński M., Jarmołowicz-Nowikow E. (2010), *Prosodic and Gestural Features of Phrase-internal Disfluencies in Polish Spontaneous Utterances*, Proceedings of Speech Prosody 2010 Conference, Chicago.

43. Keller E. (2005), *The analysis of voice quality in speech processing*, [in:] *Lecture Notes in Computer Science*, Chollet G., Esposito A., Faundez-Zanuy M. [Eds.], **3445**, pp. 54–73, Springer-Verlag.

44. Ladd D., Silverman K. A., Tolkmitt F., Bergmann G., Scherer K. R. (1985), *Evidence for the independent function of intonation contour type, voice quality, and f0 range in signalling speaker affect*, Journal of the Acoustical Society of America, **78**, 2, 435–444.

45. Laver J. (1980), *The Phonetic Description of Voice Quality*, Cambridge University Press, Cambridge.

46. Local J., Kelly J. (1980), *Projection and 'silences': Notes on phonetic and conversational structure*, Human Studies, **9**, 185–204.

47. Ludlow C. L., Coulter D. C., Bassich C. J. (1982), *Relationships between vocal jitter, age, sex, and smoking*, Journal of Acoustic Society of America, **71**, 55–56.

48. Maclay H., Osgood C. E. (1959), *Hesitation Phenomena in Spontaneous English Speech*, Word, **1**, 19–43.

49. Mary L., Yegnanarayana B. (2008), *Extraction and representation of prosodic features for language and speaker recognition*, Speech Communication, **50**, 782–796.

50. Nishinuma Y., Hayashi A. (2004), *Silent Pauses in Simulated Request-Refusal Type Dialogues. A Phonetic Analysis of German, Korean, and Japanese*, Symposium of Nordic Association for Japanese and Korean Studies (NAJAKS), Göteborg, Sweden.

51. Ogden R. (2001), *Turn transition, creak and glottal stop in Finnish talk-in-interaction*, Journal of the International Phonetic Association, **31**, 139–152.

52. Pakosz M. (1982), *Intonation and attitude*, Lingua, **56**, 153–178.

53. Patel S., Scherer K. R., Björkner E., Sundberg J. (2011), *Mapping emotions into acoustic space: The role of voice production*, Biological Psychology, **87**, 93–98.

54. Rosner B. S., Pickering J. B. (1994), *Vowel perception and production*, Oxford University Press, Oxford.

55. Rusz J., Cmejla R., Ruzickova H., Ruzicka E. (2011), *Quantitative acoustic measurements for characterisation of speech and voice disorders in early untreated Parkinson's disease*, Journal of Acoustic Society of America, **129**, 1, 350–367.

56. Saks M. J., Koehler J. (2005), *The coming paradigm shift in forensic identification science*, Science, **309**, 5736, 892–895.

57. Saville-Troike M. (1985), *The place of silence in an integrated theory of communication*, [in:] *Perspectives on silence*, Tannen D., Saville-Troike M. [Eds.], Norwood, NJ, Ablex.

58. Scherer K. R., Banse R., Wallbott H. (2001) *Emotion inferences from vocal expression correlate across languages and cultures*, Journal of Cross-Cultural Psychology, **32**, 76–92.

59. Schötz S. (2006), *Prosodic Cues in Human and Machine Estimation of Female and Male Speaker Age*, [in:] *Nordic Prosody: Proceedings of the IX-th Conference*, Bruce G., Horne M. [Eds.], pp. 215–223, Peter Lang Publishing, Lund.

60. Shriberg E. (1994), *Preliminaries to a Theory of Speech Disfluencies*, PhD Thesis, Dep. of Psychology, University of California, Berkeley.

61. Shriberg E. (1999), *Phonetic consequences of speech disfluency*, Proceedings of International Congress of Phonetic Sciences, pp. 619–622, San Francisco.

62. Shriberg E. (2001), *To "Errrr" is Human: Ecology and Acoustics of Speech Disfluencies*, Journal of the International Phonetic Association, **31**, 1, 153–169.

63. Shriberg E. (2007), *High-level Features in Speaker Recognition*, [in:] *Speaker Classification I*, Mueller C. [Ed.], pp. 241–259, Springer-Verlag, Berlin – Heidelberg.

64. Stepanova S. (2007), *Some features of filled hesitation pauses in spontaneous Russian*, Proceedings of ICPhS XVI, pp. 1325–1328, Saarbruecken.

65. Swerts M. (1998), *Filled pauses as markers of discourse structure*, Journal of Pragmatics, **30**, 485–496.

66. Trager G. L. (1960), *Taos III, paralanguage*, Anthropological Linguistics, **2**, 2, 24–30.

67. Trager G. L. (1961), *The typology of paralanguage*, Anthropological Linguistics, **3**, 1, 17–21.

68. Trager G. L. (1964), *Paralanguage: A first approximation*, [in:] *Language in culture and society*, Dell Hymes [Ed.], pp. 274–288, Harper and Row, New York.

69. Trask R. L. (1996), *A Dictionary of Phonetics and Phonology*, Routledge, London.

70. Vasilescu I., Adda-Decker M. (2007), *A cross-language study of acoustic and prosodic characteristics of vocalic hesitations*, [in:] *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue*, Esposito A., Bratanić M., Keller E., Marinaro M. [Eds.], pp. 140–148, IOS Press.

71. Vasilescu I., Candea M., Adda-Decker M. (2004), *Hésitations autonomes dans 8 langues: une étude acoustique et perceptive*, Workshop MIDL04, Paris, France.

72. Vasilescu I., Candea M., Adda-Decker M. (2005), *Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages*, Proceedings of InterSpeech 2005, pp. 1773–1776, Lisbon, Portugal.

73. Wallbott H. G., Scherer R. S. (1979), *Normal speech – normal people. Speculations on paralinguistic features, arousal, and social competence attribution*, Proceedings of the Social Psychology and Language Conference, Bristol.

74. Ward N. (2004), *Pragmatic Functions of Prosodic Features in Non-Lexical Utterances*, Proceedings of Speech Prosody 2004 Conference, pp. 325–328, Nara, Japan.

75. Wu Ch.-H., Yan G.-L. (2004), *Acoustic Feature Analysis and Discriminative Modeling of Filled Pauses for Spontaneous Speech Recognition*, Journal of VLSI Signal Processing, **36**, 91–104.