Utterance segmentation and turn-taking in spoken dialogue systems

Jens Edlund[♥], Mattias Heldner[♥♣] & Joakim Gustafson[♣]

♥ Centre for speech technology (CTT), KTH, Stockholm, Sweden

♣ Voice Technologies, Expert Functions, Teliasonera, Haninge, Sweden

{edlund,mattias}@speech.kth.se, joakim.gustafson@teliasonera.se

Utterance segmentation and turn-taking in spoken dialogue systems

Jens Edlund, Mattias Heldner, & Joakim Gustafson

A widely used method for finding places to take turn in spoken dialogue systems is to assume that an utterance ends where the user ceases to speak. Such endpoint detection normally triggers on a certain amount of silence, or non-speech. However, spontaneous speech frequently contains silent pauses *inside* sentence-like units, for example when the speaker hesitates. This paper presents `/nailon/`, an on-line, real-time prosodic analysis tool, and a number of experiments in which end-point detection has been augmented with prosodic analysis in order to segment the speech signal into what humans intuitively perceive as utterance-like units.

Introduction

The most commonly used method for finding suitable places to take the turn in current spoken dialogue systems is to assume that an utterance ends where the user ceases to speak. In commercial and research dialogue systems, such endpoint detection normally triggers on a certain amount of silence, or non-speech. The method makes sense; given that a speaker is allowed to complete what she/he intends to say, the end of the utterance is likely to coincide with silence. The method yields reasonably sized units in many cases, often corresponding to sentences or some sentence-like units.

However, spontaneous speech frequently contains silent pauses inside what we would intuitively group into sentence-like units, and inside what are indeed semantically coherent units. Typical examples are topicalisations, such as "Hagagatan 14 <long silence> when was the apartment built?", and hesitations such as "I am standing to the left of a <long silence> brown building" and "take <long silence> this lamp".

There are a number of common tasks within speech technology and natural language processing where it would be useful to perform automatic utterance chunking in a way that better matches what humans perceive as utterances. The

following two groups summarise our primary motivation for this work, but other applications may apply.

- Spoken language understanding, topic detection, information retrieval, etc. Interpretation of utterances ought to be helped by sensible utterance segmentation. For example, utterance unit segmentation may make the 'speech understanding' task easier by reducing the search space in the automatic speech recognition and semantic interpretation (or parsing) components (e.g. Batliner, Buckow, Niemann, Nöth, & Warnke, 2000). In fact, the perplexity of a language model is likely to be lower if the utterances used to build the model are segmented in a consistent manner.

- Interaction control, turn-taking, back-channeling, system barge-in. Utterance segmentation is essential in the *conversational* components of a dialogue system, notably for identifying suitable places to speak (Heldner, Edlund, & Carlson, forthcoming). A dialogue system is likely to be perceived as more natural if it has a better idea of when human interlocutors have finished talking.

In advanced spoken dialogue systems, spoken language understanding and interaction control are combined. The AdApt system, for example, uses a semantically based approach in order to deal with the problems that occur as a result of silence-only based utterance segmentation (Bell, Boye, & Gustafson, 2001). Another semantic approach is used in (Skantze & Edlund, 2004).

Background

Although there is no consensus as to exactly what defines an utterance unit, we will use utterance units much in the spirit of Traum & Heeman (1997). That is, as stretches of speech bounded by prosodic cues, such as boundary tones and silent pauses. Or put differently, as speech delimited by prosodic boundaries. We also share the opinion of Traum & Heeman (1997) that the best units for dialogue systems are the very same ones that humans use. Humans use prosodic boundaries to delimit speech units both when speaking and listening. This structuring reflects the speakers' internal organisation of the information, and facilitates the listeners' processing of the message (e.g. Heldner & Megyesi, 2003).

Previous work has shown that the acoustic signalling of prosodic boundaries is a complex one, and a number of acoustic correlates of prosodic boundaries, including speaking rate, intonation, intensity and voice quality phenomena have been proposed. Among those, boundary tones and silent pauses are generally held to be the most important ones, with the duration of the silent pause being positively correlated with the rank of the boundary (Fant & Kruckenberg, 1989).

The awareness that silent pauses are important for signalling prosodic boundaries is made use of in techniques for chunking the speech stream into manageable units for speech technology applications. The end-of-utterance (EOU) detectors in state-of-the-art automatic speech recognition typically rely exclusively on a silence threshold somewhere between 500 and 2000 ms for delimiting the units to be recognised (cf. Ferrer, Shriberg, & Stolcke, 2002 and references mentioned therein). That is to say that the output of the recogniser comes in chunks corresponding to speech bounded by 'long enough' silent pauses.

However, as noted above, spontaneous speech frequently contains silent pauses also within segments we would intuitively call utterance units, and within segments that are indeed semantically coherent units. Human listeners can discriminate these utterance-internal pauses from utterance-final ones using prosodic and gestural cues, but these pauses are often well above the silence thresholds in the end-of-utterance detectors. Moreover, these silences often occur before semantically heavy words without which the unit preceding the pause may be difficult to interpret, as in the examples above.

In an attempt to resolve these problems, we have implemented a way of segmenting speech into pause bounded utterance units which substantially reduces the number of 'unfinished' utterance units – for example units ending in hesitation pauses. To do this, we use an augmented end-of-utterance detection involving standard silence detection in combination with boundary tones. Specifically, we use what we refer to as 'mid level boundary tones' to single out the utterance internal pauses from utterance final ones. These boundary tones have been observed to act as turn-keeping, or turn-holding, cues in several fairly different languages. For example, Duncan (1972) reported that any boundary tone other than the pattern with a level tone in the speaker's mid register (a 2 2 | pattern in the Trager-Smith prosodic transcription scheme) signals turn-yielding in English. Thus, the mid level boundary tone acts as a turn-keeping signal, although Duncan did not use that term. Similarly, Selting (1996) reported that level pitch accents before a pause are used to signal a turn-holding in German;

Koiso, Horiuchi, Tutiya, Ichikawa, & Den (1998) reported a strong association between a flat contour and turn-keeping in Japanese, and Noguchi & Den (1998) reported that flat intonation at the end of pause bounded phrases is an inhibitory cue for backchannels in Japanese; and Caspers (2003) reported a clear relationship between level boundary tones and turn-holding in Dutch. Although there are several observations of the function of these mid level boundary tones, to our knowledge they have never been used for utterance unit segmentation before.

Implementation

Although the preliminary tests reported in this article could be run off-line, the ultimate test for our prosodic analysis and utterance segmentation lies in on-line real-time human-computer dialogue. In order to ensure that our early results are applicable in such a situation, the implementation has to be on-line – it must not use any acoustic right context, or look-ahead. On the acoustic level, this requirement goes well with the requirements facing humans, who rarely need acoustic right context to make decisions about speech segmentation – humans on the contrary often seem to be able to predict turn endings and suchlike. Naturally, semantic expectations provide quite considerable "look-ahead" to humans, and in an ideal system they should be used in conjunction with acoustic analysis.

Furthermore, although this is not a theoretical requirement, the implementation must run in real-time, or real user studies are impossible.

The on-line real-time prosodic analysis is implemented in `/nailon/` (a phonetic anagram for online), a Tcl/Tk package based on the Snack Sound Toolkit [http://www.speech.kth.se/snack/]. `/nailon/` uses Snack to manage sounds and to extract intensity, voicing and pitch information. In its present state, the package also captures speech duration, voiced speech duration, silence duration, and the relative position of boundary tones in an online estimation of the speakers F0 range. The analysis is in some ways similar to that used by Ward & Tsukahara (2000).

A basic voice activity detection (VAD) is used to discriminate speech from non-speech (or silence). This decision is based on a noise threshold determined from the intensity distribution. A measure of the intensity (in dB) is computed for every 10 ms sound frame and the intensity distribution is updated

continuously. Any frame with more energy than the threshold is marked as speech. Although this VAD is simplistic, it is so far sufficient for our needs. VAD is a vivid research topic, but not one that will be discussed further here.

The sequence of frame-level decisions from the VAD is converted into durations of speech and silence segments with some smoothing and padding of the speech as well as the silence segments. This is done to account for various low-energy components of speech such as fricatives, short silences such as the occlusion part in stops, and various short high energy segments embedded in silences.

A pitch extractor outputs information about voiced and unvoiced speech frames, and the F0 values of the voiced frames. This sequence of frame-level voicing decisions is used to compute durations of voiced and unvoiced speech, again with some padding to account for some artefacts introduced by the pitch tracker. The F0 values in Hz are transformed into semitones relative to a fixed value. The semitone transformed F0 data are then used to estimate speaker F0 range based on the cumulative distribution of F0 data. The F0 range is bounded by a topline and a baseline defined as the cumulative mean ±2 standard deviations (also calculated cumulatively). The semitone scale is used to ensure that +2 standard deviations interval is the same musical and perceptual interval as –2 standard deviations. The F0 range is divided into three parts: high, mid and low.

The pseudo-syllabification algorithm is a modification of Mermelstein's (Mermelstein, 1975) technique to find intensity minima in the speech signal (which in turn are used to locate syllable boundaries) by calculating the difference between the intensity envelope and the convex hull of the intensity envelope. The information from the pseudo-syllabification is used together with the voicing information from the pitch extractor and the silence durations from the processed VAD data to identify voiced regions in word-final syllables (i.e. minimally the vowel) before silences in order to identify potential boundary tones. These boundary tones are classified in terms of their position in the F0 range, currently as high, mid or low tones, and in terms of their shapes: rises, falls, and level tones etc.

Finally, the speech and silent pause durations in combination with the boundary tone classification are used to make decisions about utterance unit boundaries. Any silent pause exceeding a pause threshold, following a preset minimum duration of speech, and that is not preceded by certain boundary

tones, is taken to indicate an utterance unit boundary. The thresholds and parameter values are manually set in the initial implementation used in the tests reported here. In future versions, they will be optimised based on corpus studies.

Experiments

The utterance segmentation has been tested on four different sets of speech data. The data sets are quite different, as is the available annotation for them, which has led to four different test configurations. The data used in the tests presented here were collected in different user studies at KTH and TeliaSonera.

Preliminary studies

In preparation of the implementation of a prosodic analysis capable of identifying mid level boundary tones, we performed a number of preliminary studies in which subjects were asked to produce or judge turn-takings and interruptions with regards to their naturalness. In (Heldner et al., forthcoming), the relationship between prosodic boundaries as labelled by a trained annotator and subjects' perception and production of turn-takings was explored. In a later study, two subjects were asked to divide a transcription of a five minute speech with all utterance markers (i.e. capital letters, punctuation, line breaks) removed into utterances, providing a semantic/textual utterance segmentation. This segmentation was compared with boundary tone labelling of the speech and with silence based acoustic segmentation. Both comparisons showed a high conformance. The acoustic segmentation was then manually corrected by removing any segments ending on a mid level boundary tone, with an improvement of the correspondence towards both the boundary tone labelling and the textual segmentation as a result. Several similar tests showed similar results, which inspired us to go on to implement the automatic analysis and to test it on larger data sets.

After implementation, another preliminary study was done on data from the AdApt spoken dialogue system (Gustafson et al., 2000), a Swedish-spoken multi-modal dialogue system acting as a real estate broker, providing information about apartments for sale in Stockholm. The AdApt data used comes from a user study (Edlund & Nordstrand, 2002) performed in a fully functional system where 26 subjects interacted with the system. Although the AdApt recordings contain a fair number of instances where the ASR has detected end of utterance too soon, this is often due to poor acoustic conditions.

For the preliminary test, a selection of 10 utterances annotated as prematurely segmented and 10 random utterances were chosen in order to provide a balanced sample. Automatic analysis on this sample marked 4 out of the 10 problematic utterances as unfinished, and all 10 unproblematic utterances as finished. The data is too sparse to yield significance, however.

Experiment 1

Data

The first experiment was performed on data from the CHIL project. CHIL (Computers in the Human Interaction Loop) [http://chil.server.de/] is an EU funded project aiming to "introduce computers into a loop of humans interacting with humans, rather than condemning a human to operate in a loop of computers". One of the scenarios in CHIL is a lecture where the computer may need to notify participants occasionally. The CHIL data used here was collected at University of Karlsruhe and annotated at KTH. It consists of five minutes of English speech by one single German lecturer. The language, then, is something we could call 'European English'.

Configuration

In the first experiment, the prosody enhanced segmentation was compared to segmentation produced by a pure endpoint detection algorithm. In other words, the baseline for the evaluation was an end-of-utterance detector using a silence threshold only, and this was compared to end-of-utterance detection using `/nailon/` with the VAD set to the same silence threshold. Given this configuration, `/nailon/` will produce either the same segmentation as the baseline or a smaller number of segments.

The utterance unit segmentations produced by the baseline and augmented end-of-utterance detectors were evaluated with reference to manual annotations of prosodic boundaries in the speech material. The manual annotations used a three-level convention developed within the GROG-project (Heldner & Megyesi, 2003). Each orthographic word was classified as being followed by either a weak or a strong boundary, or as not followed by any boundary. A trained phonetician annotated the entire material in three independent sessions, timed a few days apart. The majority votes of the three sessions were taken as the final classification of the prosodic boundaries. The agreement and Kappa figures for this task were 92%, and 79%, respectively. This annotation

procedure resulted in 460 words being classified as not followed by any boundary, 116 words as followed by a weak boundary, and 38 words as followed by a strong boundary. Strong boundaries were taken to indicate utterance unit boundaries; hence there were 38 utterance units in the speech material.

The following general-purpose detection metrics were used to evaluate the baseline and augmented end-of-utterance detectors: recall, precision, fallout, and error in the task of detecting utterance unit (or strong prosodic) boundaries.

Results

Whereas the baseline end-of-utterance detector segmented the speech material into 43 units (20 false alarms; 23 hits and 15 misses), the augmented end-of-utterance detector segmented it into 33 units (12 false alarms; 21 hits and 17 misses). These figures are to be compared with the 38 perceived utterance units. There was a substantial reduction in the number of false alarms when using prosodic information in combination with silence duration. The recall, precision, fallout, and error figures for the detectors in the task of detecting utterance unit boundaries are shown in Table 1.

Table 1 Recall, precision, fallout, and error figures for the baseline end-of-utterance (EOU) detector and /nailon/ in the task of detecting utterance unit boundaries

	Baseline	**/NAILON/**
Recall	0.61	0.55
Precision	0.53	0.64
Fallout	0.03	0.02
Error	0.06	0.05

Experiment 2

Data

The second experiment uses data from the HIGGINS project (Edlund, Skantze, & Carlson, 2004). HIGGINS is a spoken dialogue system project aimed primarily at investigating error handling in spoken human-computer interaction. The HIGGINS data used here consists of 20 subjects verbally describing what they see as they "walk" through a simulated city environment. The data contains a fair

number of hesitation pauses, which mainly occur when the subject is uncertain about which adjective to use to describe an object in the simulated city.

Configuration

The experiment on the HIGGINS data is a later test, using annotations specifically designed to capture whether a human listener believes an utterance is complete or not: For a number of speech segments, four separate annotators were asked to judge the likelihood of each segment being a complete utterance on a scale from one to five.

The segments were acquired by having `/nailon/` produce four five second segments ending in silence for each of the 20 speakers in the data collection. The segments were chosen randomly within the speech of each user, but with the restriction that two stimuli per speaker were judged to be hesitation pauses by `/nailon/`, and two were judged to be utterance endings proper. The actual distribution of pauses and utterance endings in the data is quite different, with a lot more endings proper than pauses, so the set of stimuli is down-sampled for an even distribution. Five of the speakers were removed because they did not produce any hesitation pauses at all, leaving 15 speakers. One speaker produced only one hesitation pause, but was left in nonetheless, making a total of 60 segments.

Results

The Z-normalised average human annotator score for the segments judged to be hesitation pauses by `/nailon/` was -0.41, and the score for the segments judged to be utterance endings proper was 0.48. Low scores indicate that the human annotators perceived the segment as unfinished. To test the hypothesis that `/nailon/` captures prosodic features that are relevant for utterance completeness, we assume that the human judges utilise prosodic (as well as semantic) features when scoring the segments. Given this assumption, there should be some covariance between the decisions made by `/nailon/` and the human judges. This was tested in a univariate ANOVA, where the decision made by `/nailon/` has a significant effect on the score made by the human judge ($F(1,228)=59$; $p < 0.05$) while the identity of the human judge does not ($F(3,228)=3$; $p > 0.05$).

Experiment 3

Data

The third test data was collected within the NICE project (Gustafson, Bell, Boye, Lindström, & Wirén, 2004), an EU funded project with the overall goal of providing users with an immersive dialogue experience in a 3D fairytale game. Spoken and multimodal dialogue is the user's primary vehicle of progressing through the story. The data collection (Boye, Gustafson, & Wiren, 2004) contains 650 utterances recorded from 10 subjects aged between 11 and 15 at the Technical Museum in Stockholm. The system was displayed on a large back-projection screen. The user could give input to the system by means of a wireless microphone headset and a wireless gyro mouse. The system was supervised, and partly controlled, from a neighbouring room.

Configuration

The experiment was done using the same method as in Experiment 2. The nature of the data is slightly different, and consists of a large number of already segmented utterances, since it is recorded in a live dialogue system. Three annotators judged each sound segment regarding the likelihood of it being a complete utterance on a scale from one to five. The annotators also discarded some segments due to technical reasons such as strong background noises and mixed voices which occurred at the beginning and end of the recording sessions, as these factors affect the pitch extraction adversely.

The remaining segments were then labelled as hesitation pauses or as end-of-utterances proper by `/nailon/`. All in all, 295 segments were labelled, out of which `/nailon/` deemed 237 to end properly and 58 to end in hesitation pauses.

Results

Under the same assumptions as in Experiment 2, a univariate ANOVA shows that the hesitation vs. end-of-utterance decision made by `/nailon/` has a significant effect on the score of the human judge ($F(1,878)=38$; $p < 0.05$) while the identity of the human judge does not ($F(2,878)=0,3$; $p > 0.05$).

Conclusions and future work

We have shown that using prosodic cues when segmenting the speech signal into utterance-like units improves creates results that are more in line with what human judges perceive to be utterance-like units. For some data, like the HIGGINS data in Experiment 2, prosody can be used to find a substantial proportion of the utterance segments produced by an end-point detector that a human judge would describe as 'wrong'. In other data, such as the child speech in the NICE data in Experiment 3, a much smaller part of the 'failed' utterance segments can be explained by the implementation of prosodic analysis we have tested here. In part, this could be explained by the fact that the NICE data contains shorter utterances, quite a few repetitions (which are by nature planned), and speech that is somewhat more command-like than the speech in the HIGGINS data. Furthermore, the acoustic analysis used was not tuned for children, and children's speech *is* somewhat different than that of grown-ups from a processing point of view. This is evident from the ASR results on the NICE data – the children had a significantly higher word error rate than grown-ups in the same domain. Nevertheless, a significant correlation exists between the acoustic features `/nailon/` extracts and human judges' perception of the completeness of these utterances.

If we are to improve spoken human-computer interaction using these techniques, for instance by making spoken dialogue systems seem more responsive and less error prone, prosodic analysis is obviously not enough. Humans use higher levels of understanding and are experts at prediction. Naturally, a perfect spoken dialogue system would want perfect ASR and speech interpretation, as well as reasoning and planning, and so on. We feel, however, that by designing a way to access prosodic information on-line, we provide one source of information that should be combined with others to guide the decisions we make about turn-taking and other tasks that require speech segmentation on the utterance level.

The results presented here are a first attempt. The next step is to fine-tune the analysis, to add more features, and to test the method on a greater variety of data. This should be followed by studies of how useful the acoustic analysis actually is in real spoken human-computer dialogue, since improving such dialogue is our primary concern. Finally, we need to merge the information provided by the prosodic analysis with other data, such as semantic interpretations, dialogue context, etc. When this is done, we hope that we will be a step closer to modelling human dialogue.

Acknowledgements

The work presented was carried out at the department of Speech, Music, and Hearing (TMH) and the Centre of Speech Technology (CTT), KTH, Stockholm, Sweden and at Teliasonera, Haninge, Sweden.

The Department of speech, music and hearing at KTH in Stockholm researches speech communication, combining work on multimodal dialogue systems and research in component speech technology with linguistics, phonetics, cognitive science, psychology, and computer science.

The Centre for speech technology is a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations.

The work was supported by the EU funded projects CHIL (IP506909), http://chil.server.de/ and NICE (IST-2001-35293), http://www.niceproject.com/.

References

Batliner, A., Buckow, J., Niemann, H., Nöth, E., & Warnke, V. (2000). The prosody module. In W. Wahlster (Ed.), Verbmobil: Foundations of speech-to-speech translation (pp. 106-121). Berlin: Springer-Verlag.

Bell, L., Boye, J., & Gustafson, J. (2001). Real-time handling of fragmented utterances. In Proceedings of NAACL 2001.

Boye, J., Gustafson, J., & Wiren, M. (2004). Deliverable D7.2b in the NICE project: Evaluation of the first NICE fairy-tale game prototype, from http://www.niceproject.com/deliverables/

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in Dutch. Journal of Phonetics, 31, 251-276.

Duncan, S., Jr. (1972). Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology, 23(2), 283-292.

Edlund, J., & Nordstrand, M. (2002). Turn-taking gestures and hourglasses in a multi-modal dialogue system. In Proceedings of the ISCA Workshop Multi-Modal Dialogue in Mobile Environments. Kloster Irsee, Germany.

Edlund, J., Skantze, G., & Carlson, R. (2004). Higgins - a spoken dialogue system for investigating error handling techniques. In Proceedings of ICSLP 2004.

Fant, G., & Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. STL-QPSR(2), 1-83.

Ferrer, L., Shriberg, E., & Stolcke, A. (2002). Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. In Proceedings ICSLP'02 (Vol. 3, pp. 2061-2064). Denver.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., et al. (2000). AdApt - a multimodal conversational dialogue system in an apartment domain. In Proceedings of ICSLP 2000 (Vol. 2, pp. 134-137). Beijing.

Gustafson, J., Bell, L., Boye, J., Lindström, A., & Wirén, M. (2004). The NICE fairy-tale game system. In Proceedings of SIGdial 04. Boston.

Heldner, M., Edlund, J., & Carlson, R. (forthcoming). Interruption impossible. In M. Horne & G. Bruce (Eds.), Proceedings of Nordic Prosody IX. Frankfurt am Main: Peter Lang.

Heldner, M., & Megyesi, B. (2003). Exploring the prosody-syntax interface in conversations. In Proceedings ICPhS 2003 (pp. 2501-2504). Barcelona.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. Language and Speech, 41(3-4), 295-321.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. Journal of the Acoustical Society of America, 58(4), 880-883.

Noguchi, H., & Den, Y. (1998). Prosody-based detection of the context of backchannel responses. In Proceedings of the 5th International Conference on Spoken Language Processsing (pp. 487-490). Sydney, Australia.

Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. Pragmatics, 6, 357-388.

Skantze, G., & Edlund, J. (2004). Robust interpretation in the Higgins spoken dialogue system. In Proceedings of Robust 2004. Norwich.

Traum, D. R., & Heeman, P. A. (1997). Utterance units in spoken dialogue. In E. Maier, M. Mast & S. LuperFoy (Eds.), Dialogue Processing in Spoken Language Systems (pp. 125-140). Heidelberg: Springer-Verlag.

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics, 32, 1177-1207.