Measuring Final Lengthening for Speaker-Change Prediction

Anna Hjalmarsson and Kornel Laskowski

KTH Speech Music and Hearing, Stockholm, Sweden

annah@speech.kth.se, kornel@cs.cmu.edu

Abstract

We explore pre-silence syllabic lengthening as a cue for next-speakership prediction in spontaneous dialogue. When estimated using a transcription-mediated procedure, lengthening is shown to reduce error rates by 25% relative to majority class guessing. Lengthening should therefore be exploited by dialogue systems. With that in mind, we evaluate an automatic measure of spectral envelope change, Mel-spectral flux (MSF), and show that its speaker-independent performance is at least as good as that of the transcription-mediated measure. Modeling MSF is likely to improve turn uptake in dialogue systems, and to benefit other applications needing an estimate of durational variability in speech.

Index Terms: end-of-turn prediction, final lengthening, rate of speech, turn-taking.

1. Introduction

Durational aspects play an important role in spoken language. Analyses of read speech reveal that syntactic boundaries are often accompanied by segmental lengthening [1]. Perceptual studies have further shown that final lengthening influences how interlocutors perceive phrase structure in read speech [2]. Final lengthening has also been observed in dialogue; analysis suggests that lengthening occurs in all phrase-final positions, but that syllable duration prior to speaker change is significantly shorter than before pauses in turn-medial position [3].

For dialogue systems, regulating the timing of dialogue contributions with the user is a crucial function. A common strategy is to interpret silences longer than a threshold as indicative of the user's end of turn. This is problematic since conversational speech is not produced at a constant pace, and pause length in turn-medial position varies [4]. Detecting cues in the speech preceding silence may therefore yield more robust end-of-turn detection, sooner. [5, 6] have shown that intonation features can be used to detect end-of-turn significantly better than a pause-only baseline.

In the current work, we focus on the role of lengthening, by itself, in predicting speaker change. We ask the question:

1. *Is final lengthening more pronounced prior to turn-medial silence than prior to silence followed by the other party's speech?*

Although [5] explored this problem, the data consisted of machine-directed speech with utterances only one syntactically complete sentence long. [3] analyzed turn-taking cues only for semantically and pragmatically "smooth switches", a subset of all speaker changes. The approach we follow is most similar to [6], using a strict operational segmentation. No data were discarded based on syntactic, semantic or pragmatic felicity. Our analysis using reference syllable duration shows that, on average, the question can be answered in the affirmative. We then ask a second question:

2. *Can final lengthening be measured accurately enough, automatically, to aid in predicting whether the consequent silence is turn-medial?*

Experiments with Mel-spectral flux (MSF) suggest that this second question can be answered in the affirmative at least as often as the first. Subsequent analysis indicates that for the proposed task of predicting turn-medial silence, MSF and syllable duration yield results which are only negligibly different.

2. Dialogue Data

The data used in this work are human-human dialogues, collected in a recording environment set up to mimic the interaction in the DEAL domain [7]. The scenario is a price negotiation between a buyer and a seller. In total, eight dialogues of about 15 minutes each were collected. They were informal, face-to-face conversations in Swedish; two female and four male subjects participated. The close-talk microphone recordings had been orthographically transcribed and the transcriptions automatically time-aligned [7] with the speech signal[1].

To analyze variation in segmental lengthening, each party's speech was segmented into inter-pausal units (IPUs). For our purposes, an IPU is a sequence of words inter-separated by silences no longer than 100 ms[2]. Our data contains 1897 such IPUs; 942 (49%) are followed by turn-medial silences, while 955 (51%) are followed by silences which precede the *other* party speaking. As in [6], we denote these two exclusive classes of IPUs as not-speaker-change (¬SC) and speaker-change (SC), respectively. The data was split to be disjoint in speakers, as shown in Table 1.

Data Set	Speaker ID(dialogues)role	# ¬SC,SC
DEVSET	♀$_B$(1-2)b, ♂$_A$(1-4)s, ♂$_D$(7-8)b	533,517
EVALSET	♀$_A$(3-4)b, ♂$_B$(5-8)s, ♂$_C$(5-6)b	409,438

Table 1: Split of IPUs in the 8 DEAL dialogues into DEVSET and EVALSET. "b" and "s" indicate the buyer and seller roles.

3. A Transcription-Mediated Measure

Perceived speech rate is a complex phenomenon that has been shown to depend on several different parameters [9]. However, our preliminary experiments suggested that unnormalized syllable duration could be appropriate for SC/¬SC classification. The available phonemic alignments were used to locate

[1] `N-align` [8] was used for forced alignment. The phone boundary timestamps were manually verified.

[2] 100 ms was used since Swedish has long plosives; were these included, plosive stops would be among our putative inter-word silences.

the boundaries of the last two syllables in each IPU. The maximum onset rule was invoked, where intra-word syllable boundaries are placed in a way that maximizes the number of consonants at the beginning of each syllable. A histogram of the durations of the last syllable in both IPU classes, derived in this way, is presented in Figure 1. Despite considerable overlap, the distributions suggest — as hypothesized — that terminal ¬SC syllables are on average longer than terminal SC syllables.

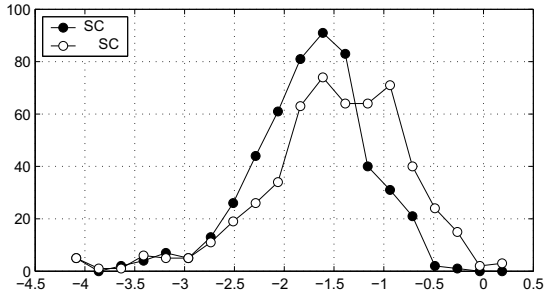

Figure 1: Unnormalized distribution of the natural logarithm of the duration (in seconds) of the last syllable in speaker-change (SC) and not-speaker-change (¬SC) IPUs.

4. A Novel Automatic Measure

As mentioned in the introduction, we seek a measure sensitive to final lengthening *differences* in SC and ¬SC IPUs.

4.1. Speech Rate Estimation without Landmark Detection

Lengthening is a local reduction in speech rate; measurement of the latter has been extensively studied. In the standard approach, a landmark detector locates salient events such as word boundaries, syllable boundaries [10, 11], syllable nuclei [12, 13, 14, 15], vowel nuclei [16], broad phone-class nuclei [17], phone boundaries [18], or phone-sequence boundaries [19]. The number of landmarks or landmark-delineated intervals is then divided by their temporal support to yield a rate[3].

Landmark detectors have been proposed which rely on low-pass-filtered energy [10], band-pass-filtered energy [12, 15], smoothed modified loudness [16, 15], zero-crossing rate [16], spectral and spectral change measures [11, 18, 13, 19, 17], spectral correlations [13, 14], temporal correlations [14], and spectral deviation from a stylized vowel spectrum [20]. They can be optimized using event detection criteria, such as the minimization of false alarms and misses. Alternately, they can be optimized to directly improve the derived speech rate estimate.

This means of estimating rate relies on how accurately one can detect landmarks, which are not in themselves of interest for the current task. Landmark localization can be argued to comprise a hard decision made prematurely. The only examples of estimation that we are aware of which avoid landmark detection are in [21], where a first spectral moment is computed over 1- to 2-second windows, and [22], where a Gaussian mixture over MFCC features is defined for each discrete rate class of interest. The former approach is simpler and therefore computationally more appealing; however, its windows are likely too long to quantify differences in final lengthening.

[3] There is also a considerable amount of literature on the segmentation of speech where a rate is *not* sought or computed.

4.2. Spectral Flux (SF)

A near-instantaneous measure of change, known as "spectral flux" (SF), or "delta spectrum magnitude", is popular in many audio processing fields. The terms appear to have been coined in [23], a year after [24] proposed "spectral dissimilarity". Several variants of this measure have found their way into the detection and/or classification of: transient notes [24], speech versus music [23], auditory scenes [25], note onsets [26], musical genres [27], and speech overlap [28]. A common aspect of these works is that they compute an inter-frame dissimilarity measure using the *complete* magnitude frequency spectrum. Maxima in the measure are thus as likely to be due to change in spectral envelope shape as to other things not of interest here, such as change in interharmonic spacing.

4.3. Mel-Spectral Flux (MSF)

To elide the effects of intonation, we measure the "flux" of the envelope only. The feature used here, Mel-spectral flux (MSF), is the logit transform of one of the "auxillary" features in [29]; its individual contribution to the task in [29] was never verified.

We compute MSF every t_{fra} seconds following standard audio pre-emphasis ($1-0.97z^{-1}$), over a centered frame whose duration is $t_{wid} = t_{sep} + 2t_{ext}$ seconds. Two Mel-frequency spectra, $\mathbf{m}_L$ for the left half and $\mathbf{m}_R$ for the right half of the frame, are computed using two windows whose maxima are t_{sep} seconds apart (cf. [29] for a precise description of the window shapes). Decoupling t_{fra} and t_{sep} enables optimization of the MSF measure for the current task.

We then compute the cosine distance between $\mathbf{m}_L$ and $\mathbf{m}_R$, which implicitly normalizes the energy in both spectra, and transform the distance using the logit function to gaussianize the resulting feature. A sign change makes it positively correlated with the notion of flux,

$$\text{MSF} \quad = \quad -\text{logit}\left(\frac{\mathbf{m}_L \cdot \mathbf{m}_R}{\sqrt{\mathbf{m}_L \cdot \mathbf{m}_L}\sqrt{\mathbf{m}_R \cdot \mathbf{m}_R}}\right) , \qquad (1)$$

where "$\cdot$" denotes the inner (dot) product. An example of the trajectory of this feature is shown in Figure 2. As can be seen, the feature — particularly when averaged over a longer interval — is high-valued during instants of high spectral change (including transitions from speech to non-speech), and lowest during the interval corresponding to the longest syllable.

The histograms of average MSF, for SC and ¬SC IPUs, in Figure 3 suggest that separation is approximately as good as for raw syllable duration in Figure 1.

5. Experiments

Binary logistic regression[4] was used to estimate the probability that an IPU is SC or ¬SC, given our transcription-mediated and/or automatic features.

5.1. Individual Features

In a first experiment, we assess the utility of the absolute duration in seconds of the last syllable in each IPU, computed as described in Section 3; we refer to this feature as "DUR1". The accuracy on DEVSET is 59.0%, with an area under the ROC curve (AUC) for SC detection of 62.7%. We have tried several transformations and normalizations of DUR1, also using the durations of the preceding syllables when available, but were

[4] As implemented in `weka` [30], version 3.5.

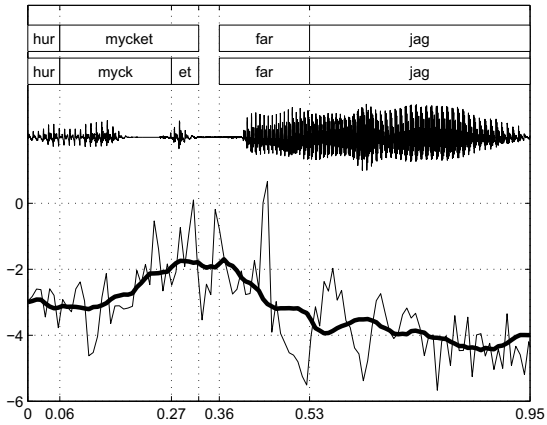

Figure 2: An 1-s IPU ("How much do I (get)?") with the word units, syllable units, audio, and MSF feature shown, from top to bottom. The MSF feature is computed every 8 ms, using frames of 32 ms with $t_{sep} = 14$ ms; its 200-ms running average, using a larger line width, is also shown. Time in seconds from IPU start, along the x-axis.

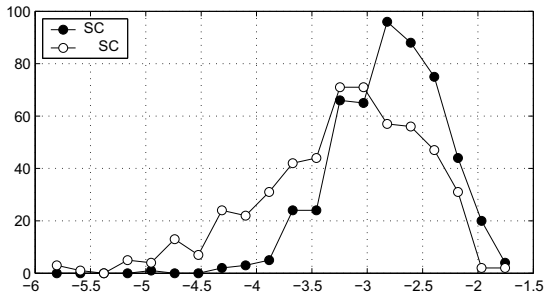

Figure 3: Unnormalized distribution of the MSF feature (computed every 8 ms using frames of 32 ms with $t_{sep} = 14$ ms) and averaged over the last 500 ms of each speaker-change (SC) and not-speaker-change (¬SC) IPU. Compare with Figure 1.

not able to beat this performance. Guessing the majority class, ¬SC, yields an accuracy of 50.8% on DEVSET.

We repeat the experiment, with the MSF feature instead of DUR1, to answer two questions. The first is the framing policy used to compute MSF; in what follows, we explore three: (1) $t_{wid} = 32$-ms frames with $t_{sep} = 14$ ms; (2) $t_{wid} = 64$-ms frames with $t_{sep} = 28$ ms; and (3) $t_{wid} = 128$-ms frames with $t_{sep} = 56$ms. In all cases, we compute the feature every $t_{fra} = 8$ ms. The second question concerns combining the time sequence of these features into a single descriptor for each IPU. We choose the average over the last τ seconds, and determine τ empirically by optimizing the AUC on DEVSET.

Figure 4 suggests that the optimal interval over which to compute the MSF average is approximately the last 400 ms of each IPU. Furthermore, it appears that computing MSF using frames which are 64 ms in duration, with two 40-ms frames whose maxima are separated by 28 ms, yields the best results. We note that using much shorter frames, even during slowly evolving syllable nuclei, can be expected to indicate high flux for creaky voice (which tends to occur at IPU end).

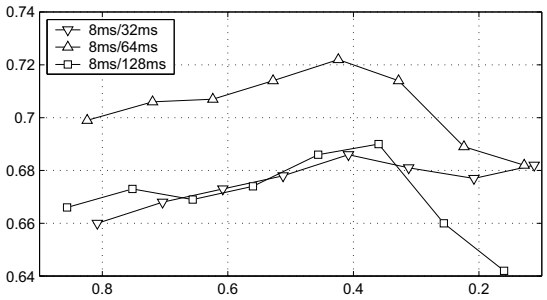

Figure 4: ROC area under the curve (AUC; along the y-axis) for DEVSET, as a function of the number of seconds (along the x-axis) of audio from IPU end, over which MSF is averaged; MSF is computed using three different framing policies.

Finally, we have also tried to model MSF dynamically; our best system involved computing a sequence of averages, in non-overlapping 150 ms intervals. The results were only negligibly better than the IPU-global average above. Combining the IPU-global averages, from all three framing policy streams, resulted in no improvement over the single 8 ms/64 ms policy stream.

5.2. Combination, Generalization and Significance

We repeat the experiments for DUR1, as well as MSF (computed every 8 ms for 64-ms frames and averaged over the last 400 ms of each IPU) on EVALSET. We also build logistic regressors for a combined 2-feature system. The results are shown in the first panel of Table 2; for DEVSET they are 10-fold cross-validation results, while for EVALSET they are single-fold results with models trained on all of DEVSET.

Data	Features	(#)	DEVSET		EVALSET	
			Acc	AUC	Acc	AUC
all	DUR1	(1)	59.0	62.7	62.5	64.7
	MSF	(1)	65.9	72.2	64.2	69.1
	⊕ DUR1	(2)	67.5	72.5	65.3	70.2
sub	DUR2	(1)	57.3	59.5	54.3	55.8
	DUR1	(1)	59.0	62.4	62.9	65.0
	⊕ DUR2	(2)	64.3	66.5	62.5	66.3
	MSF	(1)	66.2	73.0	64.9	69.7
	⊕ DUR2	(2)	66.4	73.5	66.2	70.4
	⊕ DUR1	(2)	68.0	73.2	65.5	70.6
	⊕ DUR2	(3)	68.4	73.8	67.4	71.4

Table 2: Accuracies (Acc) and area-under-the-curve (AUC) measures using different feature sets, on all of DEVSET and EVALSET, as well as on their subsets ("sub") containing at least 2 syllables.

As can be seen, on DEVSET MSF performs much better than DUR1, and their combination is negligibly better than MSF alone. On EVALSET, DUR1 is better, and MSF is worse, than for DEVSET, but their relative rank is the same.

The second panel in the table repeats the experiments, for the subset of DEVSET and EVALSET whose IPUs contain at least 2 syllables. This allows for the computation of DUR2, the duration of the penultimate IPU syllable. We observe that on both DEVSET and EVALSET, DUR1 is more informative than

DUR2 for our task; their combination helps for DEVSET but on EVALSET the results are mixed. Numerically, combining all three features appears to lead to improvement, on both data sets.

To assess the statistical significance of combining DUR1 and MSF, on all of EVALSET, we computed F-scores for DUR1, MSF, and MSF⊕DUR1. Their ranges for the detection of SC and ¬SC were 66.0–69.1% and 58.1–59.8%, respectively. A two-tailed approximate randomization test[5] revealed that the differences are not statistically significant at the $p < .05$ level. This suggests that MSF and DUR1 are measuring a very similar if not identical phenomenon.

6. Conclusions

We explored the role of IPU-terminal lengthening for speaker-independent next-speakership prediction, a function relevant to dialogue systems. Modeling the duration of the last IPU syllable, obtained using partly automatic means, yields a 25% relative reduction in classification error over guessing. We expected more errors in a fully automatic setting, since speech rate estimators are errorful and lengthening is local. Contrary to expectation, experiments show that Mel-spectral flux (MSF) offers lower error rates. We anticipate that the new feature will play a useful role in speech rate estimation, improving speech recognition, fluency diagnosis, cognitive load estimation, and others.

7. Acknowledgements

We thank Khiet Truong and Daniel Neiberg for comments. The work was supported by the Riksbankens Jubileumsfond (RJ) project P09-0064:1-E *Prosody in conversation*.

8. References

[1] Lindblom, B. and Rapp, K., "Some temporal regularities of spoken Swedish", *Papers from the Institute of Linguistics, Stockholm* **21**, 1973.

[2] Scott, D. R., "Duration as a cue to the perception of a phrase boundary", *J Acoustical Society of America* **71**(4):996–1007, 1982. doi:10.1121/1.387581

[3] Gravano, A. and Hirschberg, J., "Turn-yielding cues in task-oriented dialogue", *Proc SIGdial*, London, UK, 253–261, 2009.

[4] Heldner, M. and Edlund, J., "Pauses, gaps and overlaps in conversations", *J of Phonetics* **38**(4):555–568, 2010. doi:10.1016/j.wocn.2010.08.002

[5] Ferrer, L., Shriberg, E. and Stolcke, A., "A prosody-based approach to end-of-utterance detection that does not require speech recognition", *Proc ICASSP*, Hong Kong, China, 608–611, 2003. doi:10.1109/ICASSP.2003.1198854

[6] Laskowski, K., Edlund, J. and Heldner, M., "An instantaneous vector representation of delta pitch for speaker change prediction in conversational dialogue systems", *Proc ICASSP*, Las Vegas NV, USA, 5041–5044, 2008. doi:10.1109/ICASSP.2008.4518791

[7] Wik, P. and Hjalmarsson, A., "Embodied conversational agents in computer assisted language learning", *Speech Communication* **51**(10):1024–1037, 2009. doi:10.1016/j.specom.2009.05.006

[8] Sjölander, K., "An HMM-based system for automatic segmentation and alignment of speech", *Proc Fonetik*, 93–96, 2003.

[9] Pfitzinger, H. R., "Local speech rate perception in German speech", *Proc ICPhS*, San Fransisco CA, USA, 893–896, 1999.

[10] Mermelstein, P., "Automatic segmentation of speech into syllabic units", *J Acoustical Society of America* **58**(4):880–883, 1975. doi:10.1121/1.380738

[11] Hunt, A., "Recurrent neural networks for syllabification", *Speech Communication* **13**(3-4):323–332, 1993. doi:10.1016/0167-6393(93)90031-F

[12] Pfitzinger, H., Burger, S. and Heid, S., "Syllable detection in read and spontaneous speech", *Proc ICSLP*, Philadelphia PA, USA, **2**:1261–1264, 1996. doi:10.1109/ICSLP.1996.607838

[13] Morgan, N. and Fosler-Lussier, E., "Combining multiple estimators of speaking rate", *Proc ICASSP*, Seattle WA, USA, 729–732, 1998. doi:10.1109/ICASSP.1998.675368

[14] Wang, D. and Narayanan, S., "Robust speech rate estimation for spontaneous speech", *IEEE Trans Audio, Speech, and Language Processing* **15**(8):2190–2201. doi:10.1109/TASL.2007.905178

[15] de Jong, N. and Wempe, T., "Praat script to detect syllable nuclei and measure speech rate automatically", *Behavior Research Methods* **41**(2):385–390, 2009. doi:10.3758/BRM.41.2.385

[16] Pfau, T. and Ruske, G., "Estimating the speaking rate by vowel detection", *Proc ICASSP*, Seattle WA, USA, **2**:945–948, 1998. doi 10.1109/ICASSP.1998.675422

[17] Yuan, J. and Liberman, M., "Robust speaking rate estimation using broad phonetic class recognition", *Proc ICASSP*, Dallas TX, USA, 4222–4225, 2010. doi:10.1109/ICASSP.2010.5495686

[18] Verhasselt, J. and Martens, J.-P., "A fast and reliable rate of speech detector", *Proc ICSLP*, Philadelphia PA, USA, **4**:2258–2261, 1996. doi:10.1109/ICSLP.1996.607256

[19] Ashimura, K., Kashioka, H. and Campbell, N., "Estimating speaking rate in spontaneous speech from Z-scores of pattern distributions", *Proc INTERSPEECH*, Jeju Island, South Korea, 1433–1436, 2004.

[20] Pellegrino, F. and Andre-Obrecht, R., "Automatic language identification: An alternative approach to phonetic modelling", *Signal Processing* **80**(7):1231–1244, 2000. doi:10.1016/S0165-1684(00)00032-3

[21] Morgan, N., Fosler, E. and Mirghafori, N., "Speech recognition using on-line estimation of speaking rate", *Proc EUROSPEECH*, Rhodes, Greece, **4**:2079–2082, 1997.

[22] Faltlhauser, R., Pfau, T. and Ruske, G., "On-line speaking rate estimation using Gaussian mixture models", *Proc ICASSP*, Istanbul, Turkey, **3**:1355-1358, 2000. doi:10.1109/ICASSP.2000.861830

[23] Scheirer, E. and Slaney, M., "Construction and evaluation of a robust multifeature speech/music discriminator", *Proc ICASSP*, München, Germany, 2:1331–1334, 1997. doi:10.1109/ICASSP.1997.596192.

[24] Masri, P., *Computer modelling of sound for transformation and synthesis of musical signals*, PhD Thesis (Dept. Electrical and Electronic Engineering), University of Bristol, Bristol, UK, 1996.

[25] Peltonen, V., *Computational auditory scene recognition*, MSc Thesis (Dept. Information Technology), Tempere University of Technology, Tampere, Finland, 2001.

[26] Duxbury, C., Sandler, M. and Davies, M., "A hybrid approach to musical note onset detection", *Proc Intl Conf Digital Audio Effects*, Hamburg, Germany, 33–38, 2002.

[27] Tzanetakis, G. and Cook, P., "Musical genre classification of audio signals", *IEEE Trans Speech and Audio Processing* 10(5):293–302, 2002. doi:10.1109/TSA.2002.800560.

[28] Rossignol, S. and Pietquin, O., "Single-speaker/multi-speaker co-channel speech classification", *Proc INTERSPEECH*, Makuhari, Japan, 2322–2325, 2010.

[29] Laskowski, K., Heldner, M. and Edlund, J., "A general-purpose 32 ms prosodic vector for hidden Markov modeling", *Proc INTERSPEECH*, Brighton, UK, 2783–2786, 2009.

[30] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I., "The WEKA data mining software: An update", *ACM SIGKDD Explorations Newsletter* **11**(1):10–18, 2009.

[31] Pado, S., "User's guide to `sigf`: Significance testing by approximate randomisation", 2006. `http://www.nlpado.de/~sebastian/sigf.html`

[5] As implemented in `sigf` [31], version 2.