Predicting the attitude flow in dialogue based on multi-modal speech cues

Peter Juel Henrichsen
Copenhagen Business School
pjh.isv@cbs.dk

Jens Allwood University
of Gothenburg
jens@ling.gu.se

Abstract

We present our experiments on attitude detection based on annotated multi-modal dialogue data1. Our long-term goal is to establish a computational model able to predict the attitudinal patterns in human-human dialogue. We believe, such prediction algorithms are useful tools in the pursuit of realistic discourse behavior in conversational agents and other intelligent man-machine interfaces. The present paper deals with two important subgoals in particular: How to establish a meaningful and consistent set of annotation categories for attitude annotation, and how to relate the annotation data to the recorded data (audio and video) in computational models of attitude prediction. We present our current results including a recommended set of analytical annotation labels and a recommended setup for extracting linguistically meaningful data even from noisy audio and video signals.

Keywords: attitude detection, prediction of attitude flow, attitude annotation, multimodal speech cues

Introduction

Sharing of content and alignment of attitudes are two of the basic features and goals of human communication, most clearly in face-to-face communication. These features and goals are also present in human-computer interaction, especially when the computer is represented by an ”embodied communicative agent” (ECA). To be a natural and smooth communication partner, an ECA has to be sensitive to the attitudes of its interlocutor, thus it has to have processes for recognizing and producing attitudes. This we will call attitude administration below. We here present an analysis of the acoustic features of attitude expression in the Swedish part of the Nordic NOMCO project database.

In this paper we first discuss the challenges of attitude administration in a simplified experimental setting, viz. the prosodic component of a typical TTS system (text-to-speech). We then approach the even more difficult realm of dialogue. We believe, models of attitude administration in man-machine dialogues should build on annotated recordings of human-human conversations. We present some ideas for detecting and exploiting the correlates between the acoustic features of the speech signals and the communicated attitudes, using a subset of the Swedish NOMCO data (audio files and anvil-annotated video-files). Based on recorded naturalistic examples, we discuss how to pre-process the raw audio files and the original annotation files (ANVIL-format) preparing an automatic attitude recognizer.

Attitude administration in monologue

It is a well-established experience among constructors of synthetic voices (TTS, Text-To-Speech systems) that an incoherent or unnatural prosodic contour is extremely disturbing to the listener. Human listeners will, in general, be fairly forgiving of clumsily spliced phonetic segments and sudden clicks and cracks to the sound image; after all, we are often exposed to badly encoded speech signals in our mobile phones, and as long as the prosodic contour is authentic and the words reconstructable, we manage to compensate without too much cognitive effort. In contrast, a speech signal with a prosodic encoding out of sync with the intended message cannot be compensated by subconscious means since it is no longer redundant, but contradictory, the reconstruction effort now depending on an intellectual decision procedure. For this reason, naturally sounding prosody has a high priority in any ambitious TTS project. Unfortunately, the principles of prosody assignment are anything but simple and mechanically applicable.

Prosody is the quintessential parameter for emotions and attitudes in speech; by a subtle change in prosodic outline, an utterance may shift it's psychological effect entirely, from earnest to ironic, happy to sad, tentative to confident, or even communicating several emotions-attitudes simultaneously.

Prosody assignment, then, is ultimately an AI complete enterprise. Since genuinely intelligent reasoning systems are probably still decades away, or centuries, we currently have no better option than *mimicry*. By simulating human behavior through prosodic models trained on conversational data, at least we may be able to avoid unwanted attention traps as discussed above. Modern commercial TTS systems invariably employ large databases of human read-aloud data (usually 100+ hours). The sound repositories are, of course, aligned with phonetic transcriptions, but may also be annotated for parameters like style, mood, voice (assertive/interrogative/imperative), discourse function, and so on. By analyzing an input text through these parameters and using the result as an advanced multi-dimensional search query, a best-match for each text element is identified in the sound database. When successful, the speech produced is thus composed of played-back sound instances where the human reader was in a state matching the requirements of the text, not only phonetically, but in a generalized sense reflecting even the attitude. The best modern TTS systems often approach a 'nature identical' prosody when the input text conforms to the style and vocabulary supported in their sound database. Recent examples of TTS projects with highly conscious approaches to the psychological factors of prosody assignment include Aylett et al (2008), Oparin et al (2008), and Henrichsen (2012).

Attitude administration in dialogue

In TTS systems, a naturally-sounding prosodic rendering of an input text can often be determined through rule-based text analysis and intelligent database querying, as explained above. When entering the realm of dialogue, however, prosody assignment becomes far more challenging. Speaker A's attitude pattern must now be determined by speaker B within a very short time frame based on a wealth of multi-modal sensory data, or speaker B will be at risk of producing bizarre feedback (or other attention traps). Such attitude administration may not be perceived by humans as a great challenge, but in spoken language agents, any rules for xyz must be made explicit. Inspired by the success of data driven TTS, one could suggest to compile xyz, but no (manageable) database could ever cover the potential attitudinal variation in live conversations. What cues, then, can be computerized and exploited by an automatic agent tracking the attitude of the human interlocutor?

One approach is to build computational models trained on human-human data, applying them to recordings of dialogues. We concentrated on a sub-part of the NOMCO material consisting of eight dialogues from the "first-encounters" corpus. Our reason for selecting these eight were that (only) these conformed to these requirements:

- video+sound recording
- two extra sound tracks using high-quality chin mounted mics
- individual anvil tiers including markup for a range of attitudes (introduced shortly)
- mixed population of male and female informants

The experimental setup

All of the eight recordings contained two students meeting for the first time. Their instructions were to get to know each other. For most interactions this meant that they exchanged information about names, present occupation and interests. Both participants were standing up about 50 cm away from each other, face to face at an angle of about 90 degrees, and were filmed against a white background. They could move freely in all directions. They were typically friendly and attentive to each other.

Multi-modal corpus data - a computational challenge

As is often the case with speech signals recorded under quasi-ecological conditions, the acoustic quality leaves something to be desired with respect to signal to noise ratio, channel separation, reverberation, and so forth. In our recordings, there are several instances of over-steering (clipped samples), and the reverberation is measured to about 250dB/s corresponding to an echo of approximately 400 ms. These facts combined with a modest channel separation at 20 dB makes it difficult to

perform pitch tracking for the individual speakers (see below). Regarding the functional coding, all files were checked by a separate person than the annotator. The synchronization of the audio and video streams were out of sync by >1% in some instances and in these cases had to be manually assessed.

To these circumstantial challenges come the tractability issues. As mentioned, computational attitude prediction must be quick and responsive. CPU-heavy decoding methods are therefore not feasible (e.g. automatic speech recognition) leaving us with the 'easy', low-level acoustic parameters such as F0 (pitch), intensity, spectral tilt, and Harmonicity-to-Noise ratio (HNR). We introduce each of them in the following section. They are all very well understood in a linguistic frame of reference.

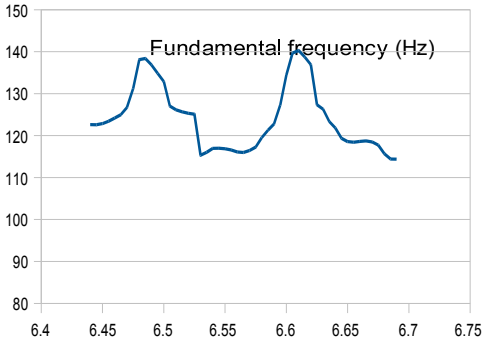

Figure 1. F0 tracing of a two-syllable word (NOMCO informant V8649L, t=6.44-6.69, utterance “eller”)

Among the acoustic features exploited by linguists, the fundamental frequency (*F0*) is probably the most popular, its interpretation in the audiological domain being so straightforward: The pitch. The difference between *pitch* and *F0* should however be noted, the former being a psychological quality and the latter, a physically well-defined2 property that can be determined by a measuring instrument independent of the human ear. A discriminating example is the so-called overtone singing. When listening to an overtone singer one experiences a succession of pitches corresponding to a certain melodic line; this line is however quite independent of the actual F0 progression and is achieved by the singer changing the filtering effect of his upper speech organs rather than changing the tension of his vocal cords. In ordinary speech, however, F0 tracings usually represent the experienced prosodic contour fairly reliably.

Fig. 1 above shows an F0 analysis of a NOMCO speech sample. The 250 ms sample represents the two-syllable Swedish word 'eller' (Eng. translation *or*) pronounced by a male speaker. This word consists of sonorants only so F0 is defined throughout. Usually, only a fraction of a speech signal will be defined for fundamental frequency since silent passages and passages without phonation (e.g. obstruents like [s] and [k]) do not produce meaningful F0 values. Observe in particular the 'wild' values, which have to be filtered away prior to the prosodic analysis. In our project, high cut-off points at 300Hz for male voices were used, 400Hz for female voices, and low cut-off points at 80Hz for both (fig.2). Even if most of the derived F0 values thus have to be abandoned as undefined or meaningless, the resulting data sparseness is not necessarily a problem for prosody analysis since the missing values can often be interpolated. Movements in the prosodic domain are, after all, relatively slow compared to the succession of phones.

Intensity is another parameter often used in acoustic-phonetic analysis. As a computational data type, this parameter has a quite different profile from F0 being *always* defined (even when the speaker is silent). In fig. 3 an intensity graph is shown for the same sound sample. Comparing the two projections it is obvious that most of the speaker's own activity is represented in the intensity range above 50dB (utterances around t=6.5”, t=11.0”, t=12.0”) while the activity of the other speaker (counting as noise in this audio channel) dominates the range 30-50dB. The limited channel separation adds to the challenges when interpreting the intensity data.

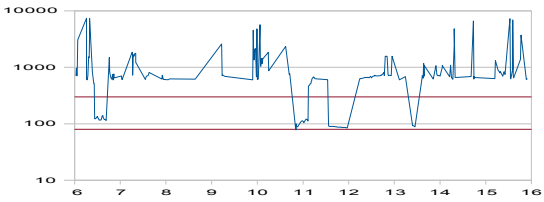

Figure 2. F0 graph, 10 seconds including the “eller” incident discussed above (t=6.45-6.70).

The two red bars indicate the filter for meaningful pitch values (80Hz<P<300Hz).

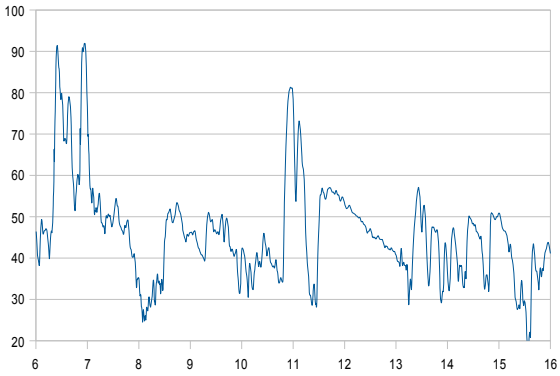

Figure 3. Intensity graph. 10 seconds' recording including the same "eller" incident as above.

The other acoustic parameters we have mentioned are both variants of the intensity parameter. *Harmonicity-to-Noise ratio* (HNR) corresponds roughly to the phonetic 'voicedness'. HNR calculation is performed by separating the harmonic components of the physical sound signal from its noise components, determining the ratio of their individual intensity (amount of energy per time unit). Language sounds with no harmonic components at all such as [s][f][h][p][t][k] and other obstruents produce very low values for HNR, due to their lack of harmonics in contrast to full vowels scoring high. The final acoustic parameter under consideration, the *spectral tilt*, may for instance be determined by comparing the intensity of a sound signal in two distinct frequency bands. Language sounds with much energy in its lower frequencies and less energy in the higher end will, under this interpretation, show a relatively large tilt. Depending on the instantiation of the filter values, various phonetic oppositions (e.g. front-back, open-close, labial-dorsal) and other features can be traced.

Acoustic parameters for prosodic analysis and attitude determination

Among the parameters we have considered, F0 is probably the most relevant for attitude detection, lending itself readily to prosodic interpretation. It should be supplemented by at least one other parameter, though, since data sparseness for F0 becomes a problem with declining acoustic quality (e.g. background noise or poor channel separation for overlapping speech). The other three candidate parameters are all robust in the sense of being defined everywhere, even for silent passages, so in a narrow sense they all serve well for data completion. However, after some initial experiments neither HNR nor spectral tilt proved suitable for our purposes. They both tend to respond more closely to the phonetic fluctuations than to the slower prosodic oscillation while of course the latter is the more important information source for attitude detection.

For these reasons we settled on a computational framework based on fundamental frequency and intensity measurements only.

The anvil annotation format for multi-modal transcription

The recordings were transcribed and annotated using the anvil annotation format for multimodal transcription (Kipp 2001). This format allows simultaneous viewing of the video recording , its transcription/annotation and listening to the audio recording. It also allows viewing of imported acoustic analysis of the audio recording from PRAAT (Boersma & Weenink 2005).

The purpose of the format is to allow analysis of different features of multimodal communicative behavior in synchronized relation to each other, e.g. the relationship of prosody to gestures and spoken words.

The annotation is done by a single annotator and then checked by another annotator. The annotators follow the GST+MSO transcription standard (Nivre 2001, 2004) and the MUMIN standard for multimodal annotation (Allwood et al. 2007).

Preparing the anvil annotations for machine learning

As mentioned, the study reported here used a sub-corpus of eight NOMCO recordings of Swedish first-encounters. The test material includes, for each encounter, one video+sound recording, two individual mono-recordings using good-quality portable microphones, and one anvil annotation tier per speaker.

The team of NOMCO annotators were, to a large degree, free to choose their own attitude labels and delimitation. As a result, the annotation material is extremely heterogeneous. The eight anvil files contain 439 reported

attitude events, the shortest lasting only a small fraction of a second (0.04”), and the longest stretching over almost three minutes (173”). The overall distribution of event durations is shown in fig. 4. Not surprisingly, the set of applied attitude labels is large and diverse: 55 English and Swedish terms, distributed over several grammatical categories.

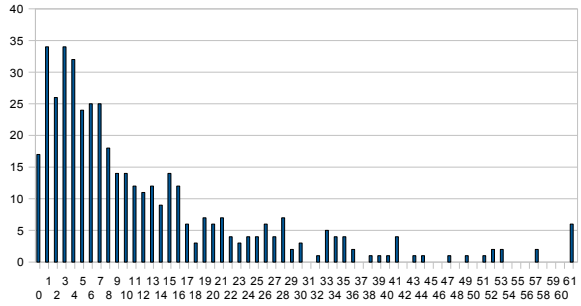

Figure 4. Distribution of attitude events as a function of their individual duration. Attitude events longer than one minute are accumulated at x=61”. Average duration = 12.7”; median = 8.0”.

A suitable subset of the attitude tags had to be extracted for machine learning purposes. As the effectiveness of learning algorithms stand or fall by the cardinality and consistency of data types represented in the training set, we excluded all sparsely used tags. In addition even some relatively densely populated attitudes (labels with many occurrences in the anvil files) had to be excluded due to the relatively low accumulated duration they represent (amount of acoustic data in terms of time frames). Since our investigations are based partly on F0 measurements, this data type being particularly fragile as discussed above, the accumulated duration for each attitude under investigation is thus at least as significant as a selection criterion as is the amount of associated events. Fig. 5 shows the set of applied attitude labels sorted by accumulated duration.

304486 Interested	124961 Friendly	121147 Casual	102014 Bored
99920 Thoughtful	65473 Confident	63998 Insight	55440 Amused
53247 Dominant	44826 Enthusiastic	42894 Uninterested	33589 Impatient
29758 Hesitant	27528 Unconfident	20371 Worried	19724 Happy
19655 Uncomfort-able	18061 Uncertain	16630 Sceptical	16518 Recogni-tion
13651	12113	9330	9210

Certain	Sad	Relaxed	Apprecia-ting
8249 Irritated	8111 Patient	7884 Nervous	7687 Surprised
7193 Describing	6688 Confused	6674 Condescending	5833 Nervous
4634 Suspicious	4553 Provocative	4290 Hopeful	2609 Sarcastic
2426 Compassion-ate	2409 Arrogant	2281 Restless	1274 Embarrasse d
1273 Questioning	1264 Explaining	1050 Confirming	1009 Amusing
967 Shy	920 Ironic	833 Convinced	433 Disappointe d
346 Understand-ing	328 Frustrated	328 Puzzled	313 Thoughtful-ness
225 Deliberative	160 Confirma-tion		

Figure 5. Applied attitude labels. The labels are sorted by accumulated duration (number of 5ms time frames).

After some formal considerations, semantic reflections, and initial experiments, we settled on a test set A10 of ten attitudes.

A10 =

{Interested, Friendly, Casual, Bored,

Thoughtful, Confident, Amused, Enthusiastic,

Uninterested, Impatient}

Each of the A10 terms is richly represented in the ANVIL files, both in terms of amount and accumulated duration. For reasons of dissemination the Swedish terms were excluded altogether (e.g. 'ifrågasättande'); also most of these were used very infrequently.

A Formal Model of Attitude Prediction

Relating the A10-based annotation data to the acoustic data based on F0/INT measurements, we arrive at the attitude profiles shown in Table 1. The profiles are based on three statistical parameters (I-III).

I. F0, standard deviation for each attitude event ('meaningful values' only, see fig. 2)3
II. INT, average for each attitude event (values relativized to the *most silent time slice* in the track)
III. INT, standard deviation for each attitude event

Average-based statistics (mean value and standard deviation) is a convenient way of minimizing the influence of irrelevant sound incidents caused by poor channel separation, echoic distortion, random acoustic events not related to the conversations, and other signal-to-noise problems. Also extreme variation in duration does not present an analytical problem in this perspective. On the flip-side, any contact is lost with the micro-structure of the attitude events when analyzing them as informational atoms, so the attitude model presented here must be a rather coarse one.4

Atti-tudes	AM	BO	CA	CO	EN
I	26.09	17.01	23.13	28.60	19.05
II	48.14	41.60	41.70	44.52	47.18
III	14.00	12.27	12.17	14.16	11.83
	FR	**IM**	**IN**	**TH**	**UN**
I	34.65	24.8	33.0	33.0	10.0
II	41.15	45.2	29.9	41.8	38.7
III	11.13	12.6	11.1	12.6	10.2

Table 1. Attitude Profiles. The A10 attitudes: AM=Amused, BO=Bored, CA=Casual, CO=Confident, EN=Enthusiastic, FR=Friendly, IM=Impatient, IN=Interested, TH=Thoughtful, UN=Uninterested.

Attitude Profiles as predictors

Each column in table 1 is interpreted as the formal profile representing the attitude in question. Consider a few examples. The A10 label 'Uninterested' is represented in the table by the vector (I, II, III) = (10.00, 38.77, 10.27), these values in turn representing a relatively low standard deviation for F0 ('little modulation') in conjunction with low values for intensity, both on average ('soft voice') and on standard deviation ('inactive articulation'). In contrast, the vector (26.09, 48.14, 14.00) for 'Amused' suggests a far more lively modulation, higher volume, and more active articulation.

Quantifying over all attitude events in the anvil files, we build a prediction table. Each event (i.e. its values for I, II, and III) selects its own attitude label among A10 as its nearest neighbor in the three-dimensional vector space. By way of example, consider the attitude event in anvil file v8649 from *t*=220.44 to *t*=224.12. Let us call it E'. This particular event – or rather, its vector – selects a label 'Enthusiastic' due to the relatively short geometrical distance between E' and 'Enthusiastic' in the three-dimensional data space spanned by I, II and III. No other attitude profile came closer to E' than 'Enthusiastic', this being the predicted attitude for E'.

We are now in a position to compare the annotated attitude for E' to the predicted attitude for the same event. In this case, the annotated and predicted attitudes were identical. Repeating this exercise for all attitude events, we arrive at the prediction table summarized in fig. 6.

Interested: Interested > Confident > Amused > Enthusiastic >> **Bored**
Friendly: Casual > Amused > Impatient > Confident >> **Bored**
Casual: Friendly > Confident > Amused > Casual >> **Thoughtful**
Bored: Uninterested > Bored > Thoughtful > Casual >> **Enthusiastic**
Thoughtful: Uninterested > Bored > Casual > Friendly >> **Confident**
Confident: Impatient > Interested > Amused > Friendly >> **Thoughtful**
Amused: Confident > Interested > Friendly > Impatient >> **Bored**
Enthusiastic: Enthusiastic > Interested > Confident > Amused >> **Bored**
Uninterested: Bored > Casual > Thoughtful > Impatient >> **Enthusiastic**
Impatient: Interested > Confident > Friendly > Casual >> **Thoughtful**

Figure 6. Attitude prediction table. Anvil labels are on the left, followed by the predicted labels sorted by geometrical distance.

The prediction table is best explained by an example. Attitude events labeled by the annotators as '*Interested*' are categorized by attitude predictor as 'Interested' (1st choice), then as 'Confident' (2nd choice), then 'Amused', et cetera, down to 'Bored' as the least likely choice. In a standard winner-takes-it-all regime, an automatic prediction algorithm would of course select the attitude minimizing the distance between the measured profile and the trained profile.

On a slightly more speculative note, one could read the interior of the prediction table as a set of 'gracefully declining' synonymy lists. Each line would then constitute a semantic theory about a particular attitude. The emerging relations between the various attitudes – 'Interested' associated with 'Amused' and 'Enthusiastic' and opposed to 'Bored', et cetera – seem to correspond fairly closely to our common sense understanding. Notice also that an intuitively weak predictor as 'casual' is also a statistically weak predictor. The suggested associations are the broadly-positive attitude qualities rather than any near-synonyms, in

contrast to the cases of e.g. 'Enthusiastic' and 'Bored' for which the suggested synonyms are much closer semantically related, and the semantic contrast to the antonyms at the other end much clearer (e.g. 'Enthusiastic' opposed to 'Bored'). In short, some generic knowledge on attitudes seems to have been transferred from the annotators to the trained model.

Conclusion

Building a conversational agent, we believe that attitude administration is indispensable. Since conversational partners are extremely sensitive to delayed or inadequate attitudinal response (e.g. showing indifference when presented with positive news, or enthusiasm when empathy was appropriate), attitude detection must be robust and effective within a short time frame. For these reasons we recommend that attitude predictions be based on acoustic measurements for F0 and Intensity for quick and robust data extraction under sub-optimal recording conditions (high-echoic and/or noisy surroundings).

An interesting off-spin of our investigation is the user-driven decision procedure in the design of the basic annotation scheme. As discussed, the annotators where allowed to select freely among all words in their vocabulary, unbiased by the academic purposes of the annotation activity. Based on our experience with the derived annotation scheme A10, we suggest this tag base for future annotation projects.

Finally, we have shown how anvil- transcribed video recordings of human-human dialogues can be used as data for training an automatic attitude detector. The trained attitude model even seemed to inherit some generic knowledge on attitudes from the human experts (the annotators) which is exactly what one hopes for in a data-driven competence model. As far as this preliminary experiment can tell, effective attitude prediction may hence be within reach even under sub-optimal recording conditions and extreme time pressure.

Acknowledgments

The research that has led to this work has been supported by the NOMCO project, which is funded by the NORDCORP program under the Nordic Research Councils for the Humanities and the Social

Sciences (NOS_HS) and the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287(SSPNet).

References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In J. C. Martin et al. (eds.) *Multimodal Corpora for Modelling Human Multimodal Behavior*. Special Issue of the International Journal of Language Resources and Evaluation. Berlin: Springer.

Aylett, M. P. & J. Yamagishi (2008) Combining Statistical Parametric Speech Synthesis and Unit- Selection for Automatic Voice Cloning; LangTech-2008, Rome.

Boersma, P., & Weenink, D. (2005). Praat: doing phonetics by computer (Version 4.3.01) [Computer program]. Retrieved from http://www.praat.org/

Henrichsen, P.J. (2012) Nature Identical Prosody; data-driven prosodic feature assignment for diphone synthesis, 4th Swedish Language Technology Conference (SLTC-2012), Lund

Kipp, M. (2001). anvil – a generic annotation tool for multimodal dialogue. In *Proceedings of Eurospeech*, pages 1367-1370.

Navarretta, C., Ahlsén, E., Allwood, J., Paggio, P. & Jokinen, K. (2011). Creating Comparable Multimodal Corpora for Nordic Languages. *Proceedings of the 18th Nordic Conference of Computational Linguistics*. Riga, Latvia, May 11-13. NEALT. pp. 153-160. See http://dspace.utlib.ee/dspace/handle/10062/16955

Nivre, J. et al. (2001). *Göteborg Transcription Standard* (GTS) 6.4. University of Gothenburg, Department of Linguistics.

Nivre J. et al. (2004). *Modified Standard Orthography* (MSO). University of Gothenburg, Department of Linguistics.

Oparin, I.; V.Kiselev; A.Talanov (2008) Large Scale Russian Hybrid Unit Selection TTS. SLTC-08. Stockh