



Feasibility of Constructing an Expressive Speech Corpus from Television Soap Opera Dialogue

Peter Rutten

VRT medialab (Flemish Radio and Television), Ghent, Belgium

peter.rutten@vrt.be

Abstract

This paper presents a study into the feasibility of extracting a corpus of expressive speech from television soap opera dialogue. We investigated how dialogue can be extracted from television production tapes, and what kind of signal quality may be expected. We analysed to what extent the scripts that are used in television production can provide a transcription of the actual dialogue. From the scripts we also estimated how much dialogue speech we can expect to find for each character. We based our analysis on 7 seasons (1145 episodes) of a soap opera produced by the Flemish broadcaster VRT. The results show that processing 100 episodes can result in 3 hours of speech for one of the main characters, or 2.5 hours of dialogue between two of the main characters. The scripts, however, do not provide a quick win for automatic annotation of the corpus - they do not provide sufficiently accurate transcriptions of the dialogue that was actually spoken by the actors.

Index Terms: expressive speech, speech synthesis, speech corpora

1. Introduction

Speech synthesis could play an important role in the development of virtual characters for drama pre-production, animated movies or the gaming industry. Super-realistic computer games, which will feature life-like human faces, are predicted to be at most two years away [1]. These systems still have to rely on natural speech recordings to be realistic - because commercial speech synthesisers are currently only capable of producing natural sounding 'neutral style' speech, and may at best provide ways to insert pre-recorded expressive cues or expressions into the output [2, 3].

More fundamental research is required before true expressive speech synthesis systems can be built. Data-driven research depends on the availability of expressive speech corpora, although it is not yet clear what the best approach to obtaining that data is (an overview can be found in [4]). One of the most commonly used strategies is a top-down approach: record actors that simulate expressive speech. This strategy gives excellent control over the contents of the database, but it can only be an approximation of the true nature of expressive speech as it occurs in everyday life. A radically different approach is to record spontaneous speech as it occurs in natural human interaction. A highly notable example is the JST/CREST ESP Project [5] which basically aimed at recording an 'almost infinite corpus' from which interesting parts can be extracted afterwards.

In this paper we investigate a compromise between both approaches, by focusing on acted dialogue recordings of a long-running television soap opera (referred to as 'soap' in the remainder of the paper). The soap format could be particularly

suited for the purpose of constructing an expressive speech corpus, because it consists mainly of dialogue. The dialogue is rich in expression, delivered by professional actors, and has gone through a strict authoring process to guarantee optimal quality and speech intelligibility. Furthermore, the scripts that were used during production may provide a valuable source of material for the annotation of the corpus.

In order to do a feasibility study, we were granted access to the production tapes and scripts of a Flemish soap that is now in its 12th season. We first investigated a single episode of the series for details of dialogue content, speech signal quality and script accuracy. Then we analysed the collection of scripts to determine how much speech we can expect to find for one single character (speaking to any other character), or for two characters (speaking to each other).

2. Television soap production

For our research we are interested in the format of a soap, the scripts and the audio production process.

2.1. Format

The main characteristics that define a soap are 'an emphasis on family life, personal relationships, sexual dramas, emotional and moral conflicts; some coverage of topical issues; set in familiar domestic interiors with occasional excursions into new locations' [6]. These characteristics, and the fact that a soap mainly consists of dialogue, make the format a potentially rich source of expressive speech. The naturalness of the expression, however, can be variable. Within the soap format there is a large variation in style, from the very melodramatic US and Latin American daytime serials, to the more realistic Australian and UK serials. It is the latter style that we would prefer as a source for an expressive speech corpus.

2.2. Scripts for soap production

A script (or screenplay) is the blueprint for the production of an episode of a soap. Script formats are not formally standardized, but within the industry several conventions are followed [7]. A script contains the following elements:

- **Synopsis:** a short description of all the scenes that make up the episode (place, characters, action).
- **Scenes:** a more detailed description of setting, location, time, characters, and expected duration for each scene.
- **Dialogue:** is described by the character's name followed by the character's speech, with optional acting directions embedded in the text.

2025.49 LOC EXT OORAMNNA CHALET
JOHN - DAG
<i>Werner, John, Caroline</i> 1'22
Avond

*WERNER ZIT OP EEN ROTS EN IS IN
GEDACHTEN VERZONKEN.
JOHN KOMT ERBIJ MET TWEE BLIKJES
BIER.*

WERNER

I ... I had lots of expectations when I came here,
but this ... (SCHUDT HOOFD)
(HAD NIET GEDACHT SOORT
VERBONDENHEID TE VOELEN MET JOHN)

JOHN

(BEGRIJPT DIT NIET ZO) When are you
leaving?

WERNER

As soon as possible. Time to go back to my life.

JOHN

(SIGH) I just want Julie to be happy, you
know...

Figure 1: *Extract from a script*

Figure 1 shows an extract from a script. It starts with a description of the scene, followed by the dialogue. The convention is to capitalize all the text that is not spoken, and to enclose acting directions in brackets.

For the purpose of creating a corpus of expressive speech, we would like to extract as much information as possible automatically from the scripts. Depending on the accuracy of the script, the character turns could be used as input to an automatic speaker recognition system, the dialogue could be aligned by a speech recogniser, and the acting directions could provide valuable information about expression.

In this paper we only used the scripts to estimate how much speech we may expect to find for each character in each episode. This helped us to select the most promising characters, how many episodes we needed to process, and what the optimal order (in terms of amount of speech) of the episodes is.

2.3. Audio production process

The sound recording and editing process in television or film production follows a more or less standard approach, as described in [8]. Dialogue is recorded on the set, where the sound director works together with a team of sound technicians to minimise noise and maximise the intelligibility of voices. Ideally, no other sound than dialogue is recorded on the set - all other sounds (footsteps, doors closing, phones ringing,...) are added in audio post-production.

The set recordings may not have the same sound quality across scenes and takes due to changing environment, use of different microphones/positions, etc. In post-production, the sound engineer will first try to create a uniform and coherent sound quality by applying several audio processing tools.

Once the dialogue tracks are ready, the sound engineer or designer will prepare a large number of tracks for atmosphere (e.g. office, traffic, restaurant background), common sound effects (e.g. cars passing, doors slamming, guns firing), synchronised sound effects (clothes rustling, footsteps, movement of hand props), design sound effects (unnatural sounds) and music.

The final step is to mix the available tracks on a system with optimal acoustics for evaluating the audio as it will be heard by audiences. The typical sound format for TV is stereo (two channels). Music tracks may have true stereo separation, but the dialogue is often identical on the left and right channel.

For foreign distribution, when the dialogue has to be dubbed in another language, it is necessary to supply a separate M&E (music and effects) mix. This M&E mix contains all the non-dialogue sound that is present on the standard stereo tracks.

It is not a common practice to archive the pre-mix dialogue tracks, so for the purpose of building a corpus of expressive speech, the dialogue will most likely have to be extracted from the production tapes. This may require advanced signal processing algorithms - e.g. to remove an M&E mix from the production mix. We did not have to go down this route, because the dialogue of the soap that we are investigating is readily available on the production tape (although mixed with some sound effects).

3. Case study - 'Thuis'

The purpose of this case study is to determine how feasible it would be to use the production tapes and scripts of the VRT soap series 'Thuis' to create an annotated corpus of expressive speech. We analysed one episode in detail to determine if we can extract 'clean' dialogue from the tapes, and how accurately the dialogue is described in the script. Then we analysed the scripts for the last 7 series (2001-2007) - to estimate how much dialogue we may expect to find for the main characters. For one of the main characters we determined the optimal sequence of episodes we should process, to get the maximum amount of dialogue for the least number of episodes.

3.1. Extracting dialogue from production tapes

Each episode of 'Thuis' is archived on a Digital Betacam tape. These tapes can store 4 channels of uncompressed 48 kHz PCM-encoded audio. For 'Thuis', the first 2 channels contain the left/right stereo sound that is broadcast. The remaining 2 channels contain the dialogue as it was recorded on the set, mixed with the synchronised sound effects that were recorded in post-production.

This means that we cannot extract a complete undistorted dialogue from the tapes. Part of the dialogue will be mixed with sound effects - making it of little use for speech research. Detailed analysis (3.2) of one episode shows that about 23% of the dialogue is mixed with sound effects, and 77% of the dialogue is clean.

To extract the audio from the tapes, we first have to convert the tapes to files with a Sony e-VTR machine. This machine produces an MXF file [9], where the audio tracks are present in 8-channel AES3 format [10]. Based on the specification of the MXF file and the AES format, we wrote a program to extract the sound as 16 bit linear PCM.

3.2. Detailed analysis of a single episode

3.2.1. Speech signal quality

The sound quality is continuously monitored on the set, to make sure that the voice recordings are as pure and intelligible as possible. This does not mean that the recorded signal is perfect - we did find a very small amount of clipping, and the loudness can show a lot of variation (about 17 dB) between scenes. We mea-

sured the signal-to-noise ratio at 20 points in the episode, across different scenes and settings, and found that it varied between 23 dB and 45 dB - with an average of 33 dB.

3.2.2. Accuracy of the script

We transcribed the dialogue text for one episode, and compared it with the script that was used during production. It turns out that the structure of the script (scenes and character turns) is accurate, but that the dialogue text itself is not accurately transcribed. The formulation of the dialogue is almost always slightly different from what is specified in the script.

When we compare the text of the script (2579 words) with the real transcription (2870 words), we find that the real transcript is 11% longer than the script. The actors tend to increase the length of the dialogue. Although the scripts are not accurate, we can still assume that the length of the script (in number of words) is a lower estimate of the real number of words that we will find in the recorded episodes.

3.2.3. Content of dialogue tracks

We used Praat [11] to annotate one episode (1400s long) for lexical speech, non-lexical speech (coughs, sighs,...), crosstalk, and sound effects (footsteps, closing doors,...). Table 1 shows that lexical speech covers 53.8 % of the episode, and non-lexical speech covers another 6%. This confirms that the soap format contains a large amount of dialogue (in this case almost 60%), making it a rich source of expressive speech.

The sound effects occur in a large part of the episode - but mostly between dialogues, e.g. when a character leaves or enters the set. In this particular episode, we found that 23% of the dialogue overlaps with sound effects or crosstalk. For further analysis we assumed that we can use about 77% of the speech on the tapes for constructing an expressive speech corpus.

Table 1: Analysis of one episode

label	duration(s)	% of episode duration
lexical	754.3	53.8
non-lexical	83.6	6.0
sound effects	370.8	26.5
crosstalk	46.5	3.3

3.2.4. Average speaking rate

The accurate transcription of the episode contains 2870 words, for 754.3 s of speech. This is an average speaking rate of 3.8 words per second. The average number of syllables per word (based on Nextens [12] transcriptions of the script) is 1.35, which leads to an average number of syllables per second of 5.1. This is slightly faster than the average speaking rate for Flemish Dutch (4.63 syllables per second), reported in [13].

3.3. Corpus size estimates

From the scripts, we estimated how much dialogue speech we may expect to find for the main characters in the soap. We then selected one character, based on how closely the character follows the standard pronunciation for Belgian Dutch and how much speech we may expect to find for that character. We then determined the optimal sequence of episodes that we should use to build our corpus, to get the most speech from the smallest number of episodes. Estimates are based on the word count in the scripts, and an average speaking rate of 3.8 words per second.

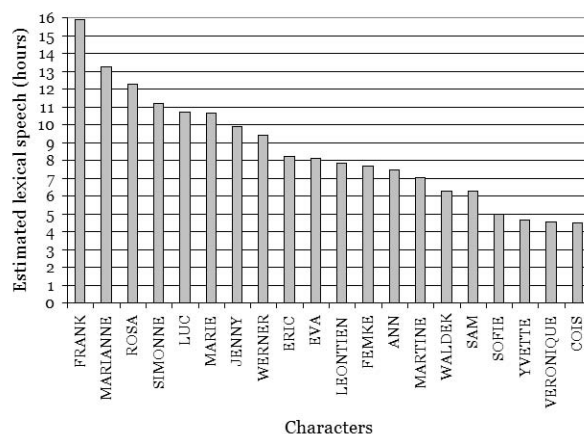


Figure 2: Estimated amount of speech in 1145 episodes, for 20 main characters

3.3.1. Selecting a character

'Thuis' belongs to the category of soaps with socially realistic storylines, focusing on everyday characters and their family life. As part of the realistic setting, most characters use some form of regional accent of Belgian Dutch. This may pose a problem if we want to use automatic analysis tools to phonetically transcribe or segment the corpus. Initially, we want to limit the scope of our project and avoid writing specialized linguistic modules. Therefore we selected the character 'Marianne' for our feasibility study. She is played by a classically trained actress, and uses a (close to) standard pronunciation.

Figure 2 shows the total amount of lexical speech we may expect to find in 1145 scripts, for the 20 characters with the most dialogue. 'Marianne' comes second in this list with 13.25 hours of expected speech.

3.3.2. Corpus size for optimised order of episodes - single speaker

The effort of constructing a corpus is directly related to the number of episodes we have to process. Each episode contributes a fixed overhead of retrieving the tape from the archive and converting it to a file (in real-time). Also, it is more convenient to annotate a large amount of dialogue for a speaker in one file than having to annotate it across a large number of files. Therefore we ordered the episodes of 'Thuis' according to the estimated amount of lexical speech for 'Marianne' in each episode. The resulting graph is displayed in figure 3. It shows that, for 'Marianne', we can expect to find 4 hours of lexical speech in 100 episodes, or about 3 hours of undistorted speech (not mixed with sound effects and crosstalk - see 3.2.3).

3.3.3. Selecting a dialogue partner

We were not only interested in a one-sided corpus of speech, we also investigated how much material we can find for two characters speaking to each other. Figure 4 shows an estimate of the amount of dialogue between 'Marianne' and the other main characters. The most productive dialogue partner, with a total of 5.5 hours of speech, is 'Ann' (Marianne's daughter in the series).

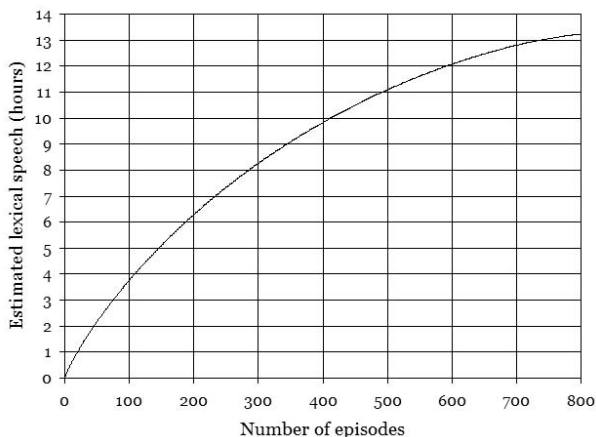


Figure 3: Estimated amount of speech for Marianne

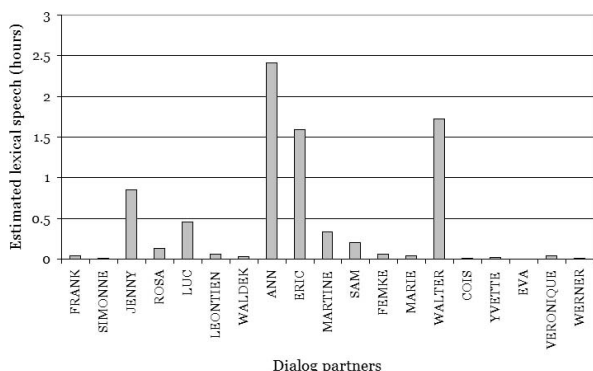


Figure 4: Estimated amount of dialogue between Marianne and 19 other main characters

3.3.4. Corpus size for optimised order of episodes - two speakers

Figure 5 shows the estimated amount of speech between 'Marianne' and 'Ann', for an optimized order of episodes. We can expect to find 3.25 hours of dialogue in the first 100 episodes, or 2.5 hours if we remove sound-effects and crosstalk.

4. Conclusion

Creating a corpus of expressive speech is a major challenge, and generally requires a large investment. In this paper we investigated the feasibility of using soap opera dialogue as a resource for expressive speech data. The case study of the VRT soap opera 'Thuis' shows that it is possible to extract speech from the production tapes, although the quality is not ideal. Part of the speech (about 23% in one episode) is mixed with sound effects, and the signal-to-noise ratio (average 33dB) varies across scenes. However, the amount of speech we may expect to find is very high - about 60% of an episode is covered by dialogue. Based on the text in the production scripts, we estimated that 100 episodes can provide 3.5 hours of undistorted speech for one of the main characters, or 2.5 hours of undistorted dialogue between two of the main characters. Although useful for estimating the amount of speech in an episode, the scripts of 'Thuis' are not accurate enough to be useful for automatic annotation of the speech.

The next step towards a corpus of expressive speech is to convert the selected tapes to file, annotate the speech of the cho-

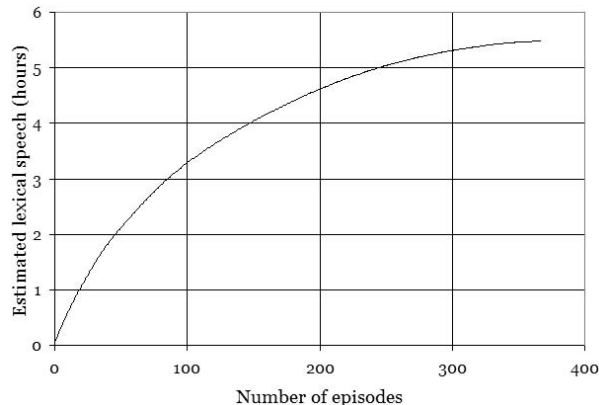


Figure 5: Estimated amount of speech between Marianne and Ann

sen character(s), transcribe their speech and annotate the distortions (sound effects and crosstalk).

5. References

- [1] BBC, "Real game characters 'next year,'" *News Technology*, Februari 2007. [Online]. Available: <http://news.bbc.co.uk/1/low/technology/6376479.stm>
- [2] P. Baggia, L. Badino, D. Bonardo, and P. Massimino, "Achieving perfect tts intelligibility," in *AVIOS Technology Symposium, SpeechTEK West*, 2006, pp. 154–164.
- [3] W. Hamza, R. Bakis, E. Eide, M. Picheny, and J. Pitrelli, "The ibm expressive speech synthesis system," in *ICSLP*, October 2004.
- [4] M. Schroder, *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. PhD thesis, Saarland University, 2004.
- [5] N. Campbell, "Databases of expressive speech," in *Proc. Oriental COCODA Workshop*, October 2003.
- [6] K. Bowles, *Soap opera: 'No end of story, ever' in The Australian TV Book (Eds. Graeme Turner and Stuart Cunningham)*. Allen & Unwin, 2000.
- [7] D. Trottier, *The Screenwriter's Bible: A complete guide to writing, formatting, and selling your script*. Silman-James press, 2005.
- [8] S. Ascher and E. Pincus, *The Filmmakers's Handbook*. Plume, 1999.
- [9] "SMPTE EG 41-2004: Material Exchange Format (MXF) Engineering Guideline." [Online]. Available: http://www.smpte.org/smpte_store/standards/
- [10] "SMPTE 331M-2004: Television - Element and Metadata Definitions for the SDTI-CP." [Online]. Available: http://www.smpte.org/smpte_store/standards/
- [11] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer." [Online]. Available: <http://www.fon.hum.uva.nl/praat/>
- [12] "Nextens: Open source text-to-speech for dutch." [Online]. Available: <http://nextens.uvt.nl/>
- [13] J. Verhoeven, G. De Pauw, and H. Kloots, "Wie Praat het Snelste Nederlands? Spreeknelheid in Vlaanderen en Nederland." *Onze Taal*, vol. 73, no. 12, pp. 336–338, 2004.