
Speaker Selection and Recordings

Keywords: *Speaker selection, recordings, normative speech, TTS*

OutsideEcho, UK
ksenia@outsideecho.com

Date	Author	Comments	Version
October 2003	Ksenia Shalonova	Request for comments and additions	0.1
May 2004	Ksenia Shalonova	Information on normative/dialectal speech included; the following sections are expanded: speech evaluation of a speaker, recordings	0.2

Contents

1. Introduction	3
2. General characteristics of a speaker	3
3. Text material to be read by a candidate speaker	3
4. Speech evaluation of a candidate speaker	3
5. Recordings	4
6. Appendix (What is Normative Speech)	4

1. Introduction

This document provides practical information for how to select a speaker for TTS and how to carry out recordings. The problems related to defining a normative (standard) pronunciation are discussed in Appendix.

2. General characteristics of a speaker

1. Age 20-40 ?
2. Female/Male – the choice may depend on the customer needs, cultural preferences etc.
3. Education – (high and preferably with the degree in philology/linguistics/literature etc.). Realises a normative standard of speech (what is normative speech see APPENDIX).
4. Place of birth (preferably the country where the TTS language is native)
5. For how long lived in the country where the motherhood TTS language is native? (at least for the last 5-10 years¹)
6. Experience in broadcast reading, theatre performances etc. (a broadcast reader or a professional actor is an ideal variant).

3. Text material to be read by a candidate speaker

Approximately 50 sentences including:

1. Sound clusters that are difficult to pronounce (e.g., consonant combinations of 3 and more sounds).
2. Different prosodic structures (yes-no question, wh-question, declarative etc.)
3. Sentences of different length

4. Speech evaluation of a candidate speaker

1. Proper realisation of all sound combinations and of different intonation types
2. Voice quality (**clear** aspirated pressed hoarse creaky constantly_clearing_of_the_throat)
3. Tempo (too slow, slow, **normal**, rapid, too rapid, varying)
4. Melodic Speech Accent (**normal** monotonous exaggerated varying)
5. Articulation (**normal** exaggerated sloppy)
6. Distinctiveness (1..5)
7. Voice pleasantness or timbre (1..5)
8. Presence/absence of non speech sounds as breath sounds, lip smacks etc. (-/+)
 - Breathing shouldn't be heard at any time.
 - For the good outcome of the recordings it is important that the speaker is able to control the amount of breath and doesn't become breathless in the middle of a sentence.

¹ to avoid accented speech

5. Recordings

Technical requirements:

1. Recording parameters: 16 bit / at least 22.05 kHz sample frequency, WAV files, mono.
2. Recording conditions:
 - Ideally: a sound-proof room to avoid noise from the computer
 - Acceptable: USB headset
3. Avoid overloading, popping (use a microphone shield)

Requirements to the speaker:

- When recording make sure that the speaker doesn't have a cold/flu.
- Have an informal talk with the speaker before making the recordings.
- The candidate shouldn't take a deep breath at the beginning of a sentence.
- Record no more than 200 sentences at a time (60-70 per hour).
- The short sentences should be read without any interruption.
- Creaky onset and offset should be avoided because they cause unusable waveforms (double pulsing). In this case there can be problems in concatenating the waveforms smoothly.

6. Appendix (What is normative speech)

TTS systems should normally generate speech that will be accepted by most local people for whom the synthesis is actually developed. For this reason speech databases for TTS are usually recorded by speakers with normative pronunciation. It is not straightforward to define what is normative speech for a language with various dialects (one dialect is normally considered to be the normative pronunciation). The standard pronunciation can be determined by several ways:

- The speech of broadcast readers of central TV/Radio stations can be considered as standard.
- Socio-linguistic study can be carried out. This type of research requires a lot of effort – plenty of recordings and their analysis. Nevertheless, it is the most reliable method as it allows verifying changes in speech culture and thus defining the normative speech (pronunciation standard typically changes significantly over a 20-30 year period).
- "Compulsory" appointment – the speech of a particular person (professor, writer, actor...) can be defined as standard.

It is important to notice that there can be 2 or even more standard variants. The choice can depend on the customer/application needs, the number of speakers etc.

It is interesting to notice that for European languages the speaker should normally have a loud and distinctive voice, whereas in Ibibio culture, for example, it is very insulting to speak loudly, so the synthesis will have to replicate a quiet voice with the corresponding voice quality (comments for Ibibio from Prof. Dafydd Gibbon).