



# Recording Speech & Music for Synthesis

---

This document assists Sensory customers with the task of recording speech that is intended for speech **synthesis**. (In contrast, customers will also want to record speech intended for voice **recognition** training. This is discussed in a separate document.) In addition, the last section of this document pertains to music files.

## Phrase list

The first step in speech synthesis design is the construction of a 'phrase list', which is a list of every utterance that the product is expected to say.

- Before attempting to write up a phrase list, you must have a clear understanding of gameplay (how the product will function and flow).
- List all sentences and phrases to be said by the product. Do not list individual words, since this sheds no light on how the words will eventually be put together (or 'concatenated') in order to form sentences. Please include translation if appropriate.
- Bear in mind that the consequences of inaccurate or incomplete scripting are time-consuming as well as costly.
- The following is an example of a phrase list for a simple record and playback device (with no voice command or notification for erase).

### Synthesis for a record & playback product

1. Record, play, or erase?
2. Recording.
3. Playback.
4. Erased.
5. Please talk louder
6. Please talk softer
7. It is too noisy here.
8. Just say: record, play, or erase.
9. Did you say "record"?
10. Did you say "play"?
11. Did you say "erase"?
12. What did you say?

- We strongly encourage customers to include the sentences given below, even if they believe at present that such sentences are not useful. If one of these phrases is not recorded, there should be a good reason for its omission.

13. It's too noisy here *or* Environment too noisy.
14. Please wait for the beep.
15. Please talk louder *or* Please talk into the microphone
16. Please talk softer.
17. Please say [*each recognition item*].
18. Just say [*each recognition item*].
19. Did you say [*each recognition item*]?
20. What did you say?
21. Recognition error / Rejected / etc.

- Depending upon gameplay and targeted market, the following are examples of other kinds of phrases that may also benefit your product:

22. I think you said [*each recognition item*]
23. *Positive feedback expressions*: good job, great, fantastic, super, wow!
24. *Negative feedback expressions*: try again, nice try, almost, sorry, that's not it, oops!

## Synthesis phrasing for successful recognition

For the recognizer to be successful, a question must be phrased so as to elicit only one *specific* answer. The answer must be specific at the level of vocabulary choice, as well as word/phrase form (singular vs. plural, definite vs. indefinite).

### Word choice

Consider the question, 'What is three plus four?' This is a good question because there is only one correct answer to this question: 'seven'. In contrast, the question, "Where does water come from?" may elicit many vocabulary items, such as "clouds," "the faucet," "rivers," "rain," and "the reservoir." Unfortunately, this question is not **vocabulary-specific**. If the speech recognizer had been coded to accept only "clouds" as the correct answer, it would appear to not work upon rejection of these other *possible* answers.

Another kind of undesirable synthesis question is one which elicits a word that varies significantly from region to region. For example, a product may ask "What do you see in the picture that begins with the letter 't'?" expecting the answer 'truck'. The recognizer will reject the response that is correct for British speakers, since 'trucks' are commonly known as 'lorries' in the UK.

### Word form

A questions must also be specific enough to elicit only a single possible form of a vocabulary item. For example, the question "What is white and puffy and gives us rain?" is specific enough to preclude

different vocabulary items like “the faucet,” “rivers,” “the reservoir,” etc., but the correct answer could be different forms of one vocabulary item: “clouds,” “a cloud,” or “the clouds.” It is therefore important that questions be phrased such that they do NOT permit variations in number (singular vs. plural) or in use of the article (“a” vs. “the”). It would be an impractical goal for the recognizer to take into account the different combinations of number and article use for each vocabulary item.

Now consider the question, “Does water come from clouds or from sunshine?” From the standpoint of word recognition, this is a good question because it targets two specific vocabulary items, each with a single form (no variation of number or article use). Both possible answers, “clouds” and “sunshine” would be programmed into the recognizer and the product would “know” the correct answer with great accuracy.

In sum, every question must be checked to ensure that there is the answer is specific at the levels of vocabulary choice and word form.

## Scripting

This section deals with the preparation of a recording session script, which is to be read by a professional voice talent.

### Sensory review of phrase list

In order to obtain the best compression possible, we suggest that you submit the phrase list to Sensory as soon as possible. Our linguistics team will review it in order to draw up a recording session script, which includes ‘concatenation breaks’. These breaks are sites within a sentence at which the voice talent will insert a slight pause so that Sensory will be able to re-use certain sounds, thereby conserving memory. For example, for phrases 9 through 11 (above), the concatenation breaks are indicated by ellipses (...), as follows:

9. Did you say... “record”?
10. Did you say... “play”?
11. Did you say... “erase”?

These breaks will permit the Sensory linguist to re-use the same utterance of ‘Did you say’ for all three questions. Without the breaks, re-use would be impossible in this case.

### Scripting preparation for foreign language products (optional)

If speech synthesis is in a language other than English, we *may* request that you send us a simple audiotape of someone reading the phrase list (Section 2, above) in a normal speaking voice, with no breaks. (This step is not necessary for all foreign languages, but will be decided on a case-by-case basis.) It is NOT necessary to have a professional voice talent read the phrase list; any native speaker will suffice. On the basis of this informal recording, the Sensory linguistics team will be able to make certain important, memory-saving decisions that relate to phrasing, intonation, re-usability, etc., and will then propose a recording session script.

## Recording

This section provides details for recording speech to be used in synthesis.

### Environment

It is best to record in a proper recording studio. In this environment, background noise and reverberation will be minimal. The result will be clear, natural-sounding speech when compressed. Furthermore, Sensory can process recordings obtained from a proper recording studio faster and more easily than recordings made in an average room since the wave forms of speech are not complicated or distorted by unwanted background noise or reverberation.

### Format

We prefer that you send us synthesis recordings in WAVE (.wav) format. Alternatively, you may send recordings on a DAT audiotape.

### Checklist of parameter settings for WAVE recordings

Record a test Wave file. Set the recording parameters for your test as described below. These settings will be saved throughout the recording session, unless you close the WAVE program. Before actually recording a test file, you must first set the sampling rate, channel, and resolution (1-3, below).

1. **Sampling rate.** Select '22.050 KHz'.
2. **Channel.** Select 'MONO'.
3. **Resolution.** Select '16 bits'.

Having set these parameters, record your test WAVE file, then check the recorded signal as outlined in steps 4 through 6, below.

4. **Saturation.** After recording a test WAVE file, check for saturation by looking at the highest points of the recordings at different scales. If the peaks of the wave forms appear flat or cut off near the top or bottom of the screen, then the recording is saturated. Adjust recording levels as appropriate in order to correct this problem. Since saturated speech does not compress well, it is imperative that you check your recordings for saturation throughout the recording session.
5. **Signal-to-noise ratio.** The proportion of speech signal to noise ought to be approximately 64 (or  $2^6$ ) to 1. Once your test file has been recorded, this ratio can be determined by checking the amplitude of the highest point of the speech signal to the highest point of the background noise level.

At Sensory, we use the following technique for establishing the signal-to-noise ratio. Find an area of the screen where you can see both the highest point of the speech signal and some background noise at the same time. Make a mental note of the height of the speech signal on the screen. Click on the vertical amplitude scale control until the height of the background noise reaches the speech signal height that you noted earlier. Each time you click is a factor of two. If you had to click the vertical scale control six times in order for the background noise level to reach the apparent height of the speech signal, then your signal-to-noise ratio is  $2^6$  (or 64) to 1.

6. **DC Offset.** On your WAVE screen, the speech signal should appear centered above and below the X-axis. If it is not, adjust the parameter 'DC Offset' in 'Tools' until the distribution of the signal is roughly equal on either side of the X-axis. A positive (+) DC Offset value will raise the signal above the X-axis; a negative (-) value lowers it. If you do not adjust the DC Offset at the outset of your recording session, then the DC Offset will have to be adjusted for each WAVE file individually.

Once you have completed steps 1 through 6, above, begin recording synthesis material. Again, continue to check for saturation and adjust recording levels as needed throughout the recording session.

### Eliciting repetitions

When we say a sentence, there are a great number of ways we can vary our 'delivery', which depends upon rhythm, intonation (melody) pattern, emphasis, emotion, attitude, personality, etc.

The voice talent should read each item on the script at least three times. This will provide us with a variety from which to select the best 'token' of a particular item on the script.

Occasionally, you may wish to have the voice talent repeat an item many more than just three times in order to attain the desired delivery. It is advisable to record more than you think you need, in case of unforeseen problems.

### Obtaining Music files from Sensory

Musical compositions for compression by Sensory must be received in four-voice General MIDI format. For these MIDI files, only certain musical (and percussion) instruments with designated note ranges are compatible with Sensory's compressed music note library. These restrictions on the acceptability of MIDI instruments and note ranges are specified in the following chart.

Acceptable Instruments & Their Ranges

Piano	D2-E7
Trombone	C3-C6
Clarinet	C3-E7
Banjo	C3-E7
Bass Drum	n/a
Snare Drum	n/a
Cowbell	n/a

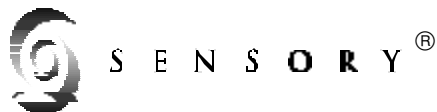
n/a = not applicable for percussion instruments

In the above notation system, C3 is two octaves below Middle C, which is C5 (550 Hz).

Another constraint on music file arrangements is that no more than four 'events' (i.e. any combination of four musical notes and/or drumbeats) may play at any one time. For example, at

time 04:01:000 in the composition, only one piano note, one trombone note, one banjo note, and one percussion beat can play simultaneously. Likewise, a piano chord of, say, three notes may be accompanied by only one other event, such as a single clarinet note.

If the fullest possible sound is desired, the composer should make the most of the four voices available. If you do not have an in-house musician who can arrange songs in order to meet the above criteria, you can send an ordinary MIDI file of the song to Sensory, and we will arrange it for you. Bear in mind, however, that the song will sound different from the arrangement you may expect, given the constraints described above.



[www.VoiceActivation.com](http://www.VoiceActivation.com)

1991 Russell Ave.  
Santa Clara, CA 95054  
TEL: (408) 327-9000  
FAX: (408) 727-4748

© 2001 SENSORY, INC.  
ALL RIGHTS RESERVED

#### **IMPORTANT NOTICE**

Reasonable efforts have been made to verify the accuracy of information contained herein, however no guarantee can be made of accuracy or applicability. Sensory reserves the right to change any specification or description contained herein.

Sensory is registered by the U.S. Patent and Trademark Office. All other trademarks or registered trademarks are the property of their respective owners.