

PART I — BEFORE YOU COLLECT ANYTHING

1. Define Why You Are Collecting Voice Data

Before recruiting anyone, answer:

1.1 What will the data be used for?

Examples:

- Automatic Speech Recognition
- Speech-to-text
- Text-to-Speech
- Speech translation
- Voice assistants
- Language documentation
- Linguistic research
- Benchmarking
- Conversational AI
- Digital government
- Healthcare
- Agriculture
- Education

Your intended use determines what type of recordings you need.

1.2 What kind of speech do you need?

There are two basic approaches used in the reference project.

a) Scripted speech

A participant reads a prepared sentence.

b) Unscripted speech

A participant responds naturally to a question or prompt.

The reference project deliberately combined the two, targeting approximately 75% unscripted and 25% scripted speech.

Practical recommendation

If your goal is a general-purpose speech dataset, consider collecting both.

Type	What it gives you
Scripted	Controlled sentences, predictable content
Unscripted	Natural speech, dialects, spontaneous phrasing
Both	Broader coverage

Do not assume that one approach automatically replaces the other.

2. Decide How Much Data You Need

Start with the target rather than starting with recruitment.

Define:

a)Target number of speakers

Example: 1,000 speakers

b)Target hours

Example: 500 hours

c)Target recordings per person

Example: 30 minutes validated speech per contributor

d)Target speech type

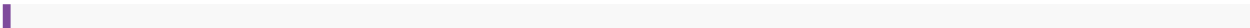
Example: 75% unscripted / 25% scripted

e)Target geographic coverage

Example: 10 counties

f) Target demographic representation

Example: Balanced across gender and five age groups.



The Africa Next Voices project targeted approximately 3,000 hours across five languages, with different targets per language.

g) Create a simple target sheet

Metric	Target
Speakers	1,000
Total hours	500
Scripted	125 hrs
Unscripted	375 hrs
Regions	10
Dialects	4
Female	50%
Male	50%
Age groups	5

Your numbers will depend on your project. The table is a planning tool, not a universal standard.

3. Define Your Language

Do not simply write:

“We are collecting Language X.”

Ask:

- Which variety?
- Which dialects?
- Which geographic areas?
- Which communities?
- Is the language internally diverse?
- Are some varieties more widely used?
- Are some varieties endangered?
- Which varieties are appropriate for the intended use?

“ Example : The AFN project explicitly considered dialect variation in languages including Kalenjin, Maasai, Somali, Dholuo and Kikuyu.

Create a language profile

Record:

1. Language:
2. Language family:
3. Country/countries:
4. Main regions:
5. Dialect groups:
6. Estimated speaker distribution:
7. Existing digital resources:
8. Existing speech datasets:
9. Native-speaker experts:
10. Relevant institutions:

4. Define Your Dialect Strategy

This is one of the most important steps.

A dataset can contain thousands of hours and still be poorly representative if most speakers come from one location or dialect.

Create a dialect matrix:

Dialect	Geography	Target speakers	Target hours
Dialect A	Region 1	200	100

Dialect B	Region 2	200	100
Dialect C	Region 3	200	100
Dialect D	Region 4	200	100

“ The Africa Next Voices project used explicit dialect targets and geographic recruitment rather than treating each language as homogeneous.

Revision #2
Created 2026-08-18 11:47:28 UTC by Angela
Updated 2026-08-18 11:57:14 UTC by Angela