Sesotho Tone Extraction - Quick Execution Guide

🚀 How to Run the Complete Pipeline

Prerequisites

Ensure your conda environment is activated:

```
conda activate sesotho-tone
```

Execution Order (In Jupyter Notebook)

Open `sesotho_tone_extraction_project/sesotho_tone_extraction.ipynb` and run cells in this order:

Phase 1: Setup & Data Exploration (Cells 1-7)

☑ **Cell 1**: Introduction (Markdown - skip)
☑ **Cell 2**: Import libraries and verify versions
☑ **Cell 3**: Set project paths (`PROJECT_ROOT` and `DATA_ROOT`)
☑ **Cell 4**: Data exploration intro (Markdown - skip)
☑ **Cell 5**: Explore audio files
☑ **Cell 6**: Analyze sample audio
☑ **Cell 7**: Load and visualize waveform

Phase 1 (Continued): Dataset Analysis (Cells 8-12)

☑ **Cell 8**: Folder structure analysis (Markdown - skip)
☑ **Cell 9**: Complete dataset structure analysis
☑ **Cell 10**: Analyze file naming patterns
☑ **Cell 11**: Data strategy recommendations

Phase 1 (Continued): Feature Extraction (Cells 13-17)

☑ **Cell 12**: Manifest generation intro (Markdown - skip)
☑ **Cell 13**: **IMPORTANT** - Create manifest & extract sample features (20 files)
☑ **Cell 14**: Robust smoke-test (5 files with error handling)
☑ **Cell 15**: Debug librosa bindings (optional - for troubleshooting)
☑ **Cell 16**: View manifest and features (optional - for verification)
☑ **Cell 17**: **OPTIONAL** - Full feature extraction (2106 files, ~10-20 min)

⚠ **DECISION POINT**:

- **Quick path (recommended for testing)**: Skip Cell 17, use sample features from Cell 13
- **Full dataset**: Run Cell 17 for all 2106 files (set `n_batches` appropriately)

Phase 2: Machine Learning Pipeline (Cells 18-26) ⭐ NEW!

- ☑ **Cell 18**: ML Pipeline intro (Markdown - skip)
- ☑ **Cell 19**: **Load and prepare dataset** - Combines features with labels
- ☑ **Cell 20**: **Speaker-independent split** - Critical for proper evaluation
- ☑ **Cell 21**: **Feature scaling** - StandardScaler on training data
- ☑ **Cell 22**: **Train Random Forest** - ~1-5 minutes depending on data size
- ☑ **Cell 23**: **Evaluate on test set** - Accuracy, F1, confusion matrix
- ☑ **Cell 24**: **Feature importance** - See which features matter most
- ☑ **Cell 25**: **Per-speaker/region analysis** - Generalization insights
- ☑ **Cell 26**: **Error analysis** - Understand misclassifications
- ☑ **Cell 27**: **Save results** - JSON, CSV, and text reports
- ☑ **Cell 28**: Summary (Markdown - skip)
- ☑ **Cell 29**: Quick reference for inference

📊 Expected Outputs

After running all cells, you should have:

Files Created

```
c:\Users\mubva\Downloads\Nlp\
├── sesotho_tone_manifest.csv          # 2106 audio files cataloged
├── features_sample.csv                 # Sample features (20 files)
├── features_sample_debug.csv           # Debug features (5 files)
├── scaler.joblib                       # Feature scaler (for inference)
├── features_parts/                     # Full extraction (if Cell 17 ran)
│   ├── features_part_001.csv
│   ├── features_part_002.csv
│   └── ...
├── models/
│   └── random_forest_baseline.joblib  # Trained model
└── reports/
    ├── baseline_results.json          # Structured results
    ├── baseline_summary.txt           # Human-readable summary
    └── test_predictions.csv           # Per-sample predictions
```

Console Output Summary

- ☑ Test Accuracy: **XX.XX%**
- ☑ Macro F1-Score: **X.XXXX**
- ☑ Training Time: **XX seconds**
- ☑ Per-speaker accuracy breakdown
- ☑ Per-region accuracy comparison
- ☑ Feature importance rankings

⏱️ Time Estimates

Phase	Cells	Time (Sample)	Time (Full Dataset)
Setup & Exploration	1-11	~2 minutes	~2 minutes
Feature Extraction	12-17	~2 minutes	~15-30 minutes
ML Pipeline	18-29	~5 minutes	~5-10 minutes
TOTAL	**1-29**	**~10 minutes**	**~20-45 minutes**

🎯 Quick Start (Minimum Viable)

If you want results FAST (for testing):

1. **Run Cells 1-16** (skip Cell 17) - Uses 20-sample features
2. **Run Cells 18-29** - Complete ML pipeline
3. **Check** `reports/baseline_summary.txt` for results

Total time: **~10 minutes**

🔧 Troubleshooting

Issue: "No feature files found"

Solution: Run Cell 13 (sample extraction) or Cell 17 (full extraction) first

Issue: "Manifest not found"

Solution: Run Cell 13 which creates the manifest

Issue: librosa keyword argument error

Solution: Already fixed! All librosa calls use keyword arguments (e.g., `y=y, sr=sr`)

Issue: All-NaN F0 values

Solution: Already handled! The `estimate_f0` function has 2-tier fallback (pyin → piptrack)

Issue: Memory error during full extraction

Solution: Reduce `BATCH_SIZE` in Cell 17 from 200 to 100

📈 Performance Benchmarks

Based on `.github/copilot-instructions.md` thresholds:

Level	Test Accuracy	Macro F1	Status

Level	Test Accuracy	Macro F1	Status
Excellence	≥90%	≥0.85	Publishable quality
Target	≥85%	≥0.80	Strong academic result ☑
Minimum	≥70%	≥0.65	Project completion

Your goal: **≥85% accuracy for strong academic submission**

🚀 Next Steps After Baseline

If you have time remaining (check date: deadline is Oct 24, 2025):

Priority 1: LSTM Sequence Model (1-2 days)

- Extract frame-level F0 contours instead of aggregated stats
- Train LSTM to model pitch trajectory
- Expected improvement: +5-10% accuracy

Priority 2: Hyperparameter Tuning (few hours)

- Use `GridSearchCV` or `RandomizedSearchCV`
- Optimize `n_estimators`, `max_depth`, `min_samples_split`
- Expected improvement: +2-5% accuracy

Priority 3: Feature Engineering (few hours)

- Add F0 contour slope, range, variance
- Add formant frequencies (F1, F2)
- Add delta and delta-delta features
- Expected improvement: +3-7% accuracy

💾 Saving Your Work

Before closing Jupyter:

1. **Save notebook**: `Ctrl+S` or File → Save
2. **Export results**: Already done automatically in Cell 27
3. **Backup models**: Copy `models/` and `reports/` to safe location

📚 Additional Resources

- **Detailed Guide**: `sesotho_tone_extraction_project/README.md`
- **Project Plan**: `sesotho_tone_extraction_project/PROJECT_ROADMAP.md`
- **AI Agent Guide**: `.github/copilot-instructions.md`
- **Feature Extraction Script**: `run_full_extraction.py` (alternative to Cell 17)

☑ Success Checklist

- ☐ Environment activated (`conda activate sesotho-tone`)
- ☐ All libraries imported successfully (Cell 2)
- ☐ Manifest created (Cell 13)
- ☐ Features extracted (Cell 13 or 17)
- ☐ Model trained (Cell 22)
- ☐ Results saved (Cell 27)
- ☐ Accuracy meets minimum threshold (≥70%)
- ☐ Speaker-independent evaluation verified (Cell 20)
- ☐ Reports generated in `reports/` directory

Ready to submit! 🎓