

A SYLLABLE BASED APPROACH FOR IMPROVED RECOGNITION OF SPOKEN NAMES



Abhinav Sethy

Shrikanth Narayanan

S. Parthasarthy



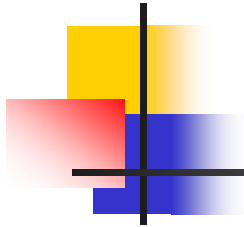
Motivation

- Performance of traditional speech recognition systems is poor for spoken names
- Direct application areas for spoken name recognition include directory assistance, travel reservations systems, banking etc
- Good performance in spoken name recognition is critical for wide acceptance of LVCSR systems

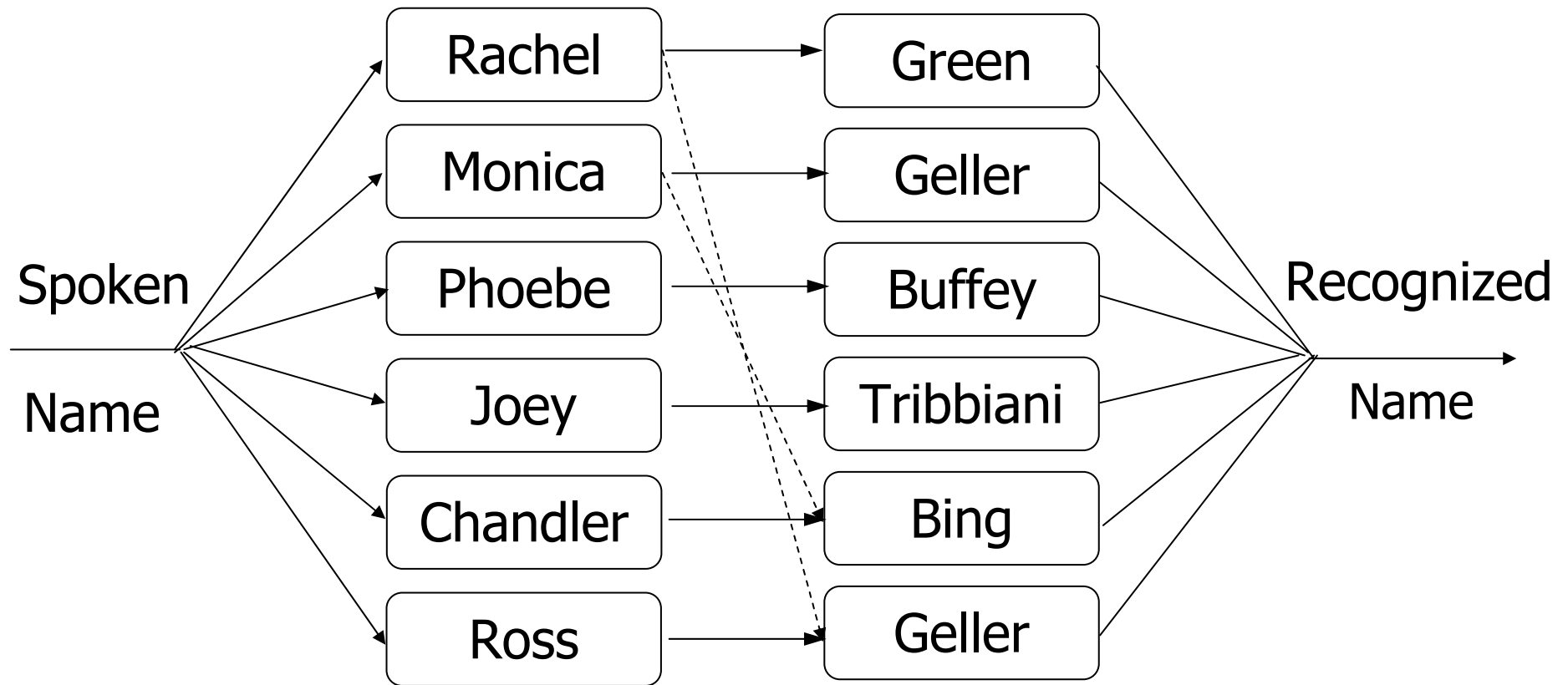


Spoken Name Recognition: Issues

- Names have a complex linguistic nature
- Pronunciation differs widely based on speaker and name nativity
- Lack of authoritative pronunciation dictionaries
 - `Each person is an authority for his own name and a novice for other names'
- Name lists for recognition can be prohibitively large. Public directories contain more than a million names!



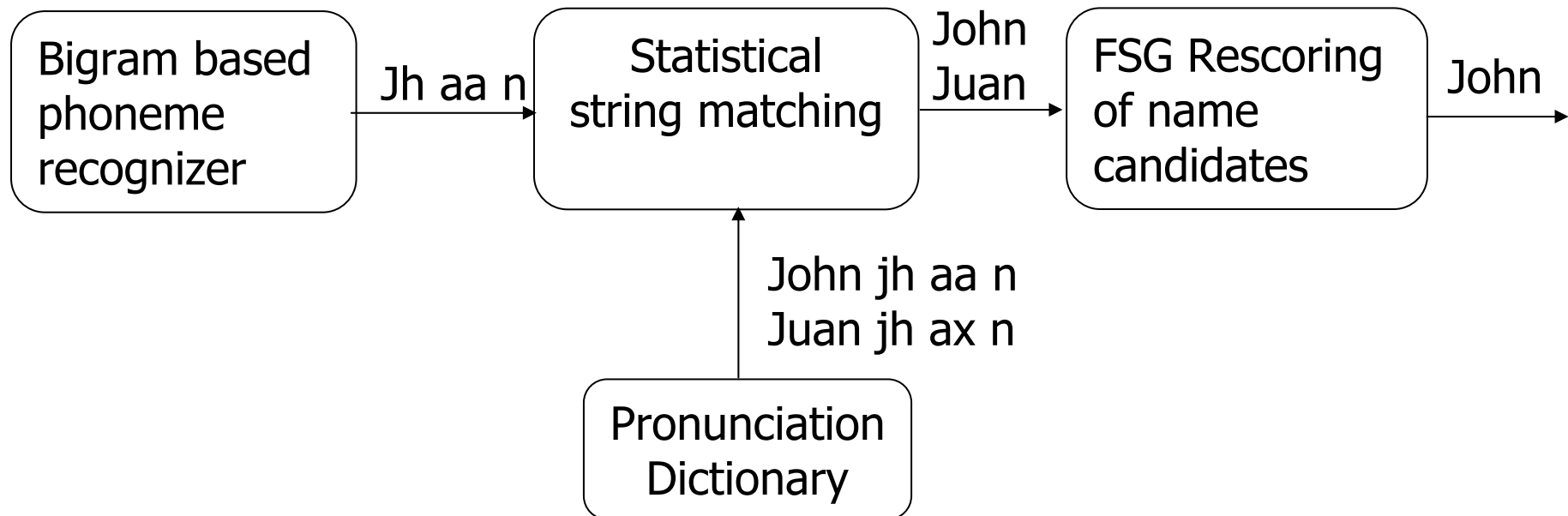
Recognition System Design: FSG



Recognition network consists of all names
Can be prohibitive for very large name lists(~100K)
Requires multiple phone level pronunciations

Recognition System Design

Reverse Retrieval



- Scalable: Can handle very large names lists
- Performance: Comparable to FSG for large name lists
- Efficient: Prunes search space using the bigram recognizer
- Issues: Requires multiple phone level pronunciations
Bigram loop based recognizer has low accuracy



Syllable Based Name Recognition Motivation

- More robust against changes in pronunciation because of extended context information
- Closely tied with human perception of speech
- Low addition and deletion rate
- Effectively incorporate suprasegmental speech phenomenon
- More suitable for reverse retrieval because of high recognition accuracy of base sequence

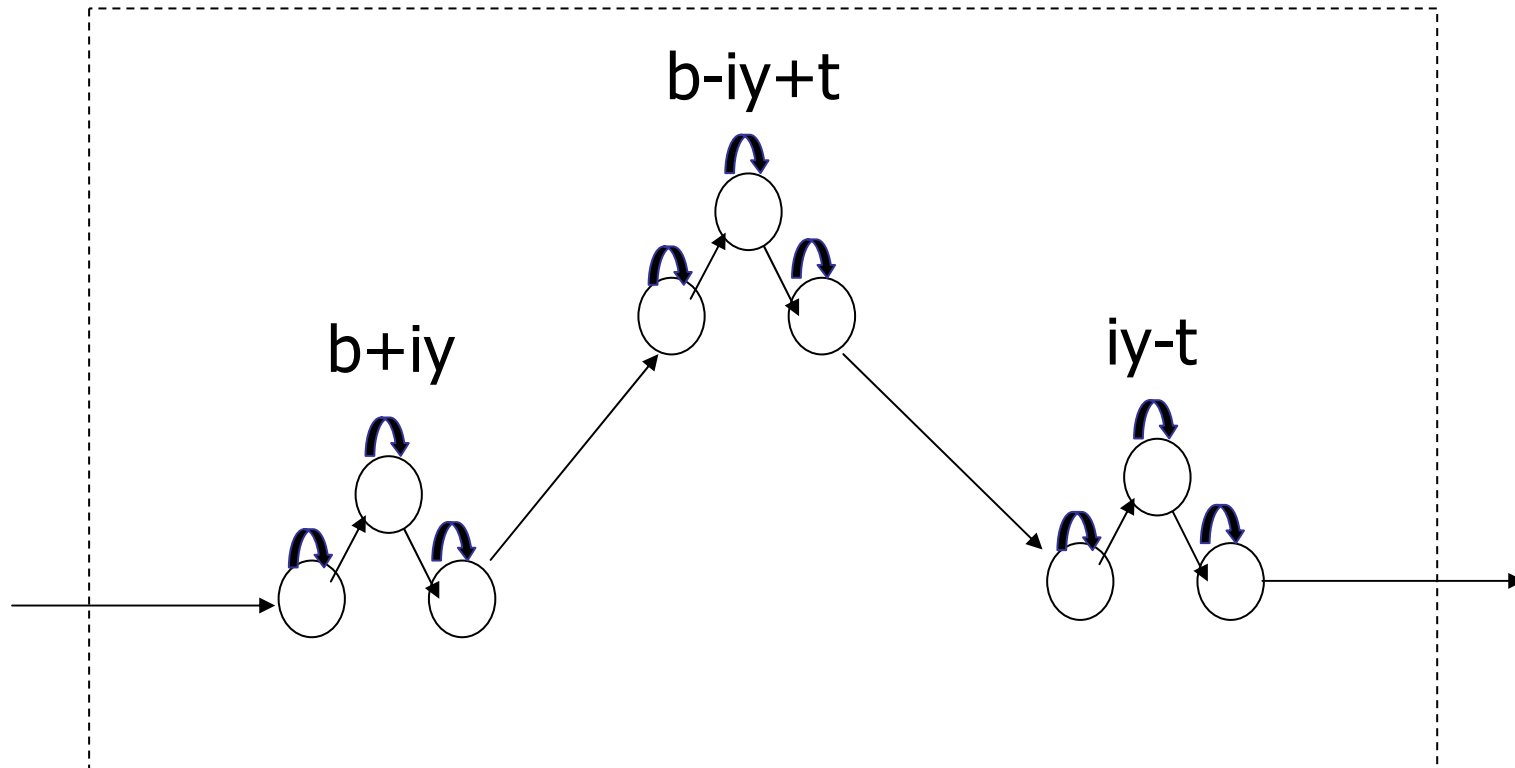


Syllable Based Name Recognition Design

- Lexical symbols are based on underlying phone sequence
 - ‘And’ with transcription `ax n d` becomes ax_n_d
- Syllabification was done using software from NIST
- Number of states for syllable model was taken to be a multiple of number of phonemes
- Syllable models had more number of mixture components to account for higher contextual variance
- No context information was used for syllable models

Syllable Based Name Recognition Design

- Syllable models are initialized from CD phone models



Initialization of the 9 state syllable b_{iy_t}



Syllable Based Name Recognition System Design

- All syllables in the lexicon will not have adequate coverage in training data
- Two systems are proposed to address the insufficient training data problem
- System I: Pure syllable recognizer
 - Create models for all syllables in lexicon using CD Phone based initialization
 - Models for which acoustic training data is lacking are not updated



Syllable Based Name Recognition System Design

- Scheme 2 : Hybrid Recognizer
 - Syllables which have good coverage in training data are used along with phoneme models
 - Context for phoneme models is based on the left and right syllable or phone models

Example

Andrew : `ax n d r uw' has syllables ax_n_d and r_uw. ax_n_d (and) is a common syllable.

Syllable unit ax_n_d is retained

Phone level units are instead of for r_uw

Thus Andrew is represented as

`ax_n_d d-r+uw r-uw'



Training: Names Corpus

- OGI Names corpus is a collection of name utterances including first, last and full names
- Speech data was collected over a telephone
- Name pronunciation is natural as the names are not read off a list
- About 10000 unique names in the corpus
- Covers 40% phonetic bigram contexts possible
- Utterances which are phonetically labeled were used to make a names dictionary



Training

- Training was done in two stages
- Phoneme, Pure Syllable and Hybrid recognizers were built for TIMIT
- Phoneme models for NAMES were initialized from TIMIT
- Syllables common to both TIMIT and NAMES were initialized from TIMIT if they had adequate coverage



Results : FSG

Recognizer Type	Recognition accuracy
Context Independent Phone	45
Context Dependent Phone	63
Context Independent Hybrid	75
Context Independent Syllable	80

Table 1: Recognition rate for different FSG based spoken name recognition systems



Results : Reverse Lookup

Acoustic Unit	Recognition accuracy after FSG rescoring
Context Independent Phone	45
Context Dependent Phone	61
Context Independent Phone	73

Table 2: Spoken name recognition accuracy for the Information retrieval scheme after FSG rescoring of Compacted name list



Conclusion

- Our experiments show that the syllable is a promising acoustic unit for spoken name recognition
- Using syllables improves accuracy for FSG networks
- Reduces the need for multiple pronunciation dictionaries
- Reverse lookup using syllables can help reduce system complexity substantially with a small tradeoff in accuracy



Future Directions

- Investigate ways of addressing the training sparsity problem for syllables
- Incorporate knowledge of phone add/del/ins to get a better distance criteria for reverse lookup
- Study the effect of name origin on recognition accuracy and automatic pronunciation generation