Finite State Automata Approach for Text to Speech Translation System in Indonesian-Madurese Language

Fika Hastarita Rachman, Qudsiyah, Firdaus Solihin
Informatics Department, University of Trunojoyo Madura, East Java, Indonesia
fika.rachman@trunojoyo.ac.id

Abstract. Madurese language is one of the regional languages in Indonesia. This language used by the Madurese people. Preservation of Madurese language now is minimal. Many migrants come from outside Madura, so communication between the surrounding communities in the Madura region often uses the national language, Indonesian. The use of Madurese as a language of communication began to decrease. This research is an effort to preserve Madurese language by utilizing the translator system technology. Madurese language is a regional language that is difficult to learn, because there are many differences found between writing and pronunciation. To overcome this problem this reasearch develop text to speech module in the Indonesian-Madurase language translation system. There are 3 versions of Madurese language level: Enja'-iyyeh, engghi-enten, and enggi-bunten. The conversion of text into sound is used with the help of syllable recording data created by the author. The process of chopping words into syllables is done using the two-level Finite State Automata (FSA) method. The output of the first level FSA becomes input for the second level. The application of FSA in Text to Speech applications is effectively used with accuracy value of 90%. The resulting sound output is in accordance with the results of syllables, but the pronunciation of some translated sentences does not have the correct intonation. The accuracy results of the intonation pattern in pronunciation of the system is 85%.

Keyword: Madurese language, translation system, text to speech, Finite State Automata, Free Context Parsing algorithm

1. Introduction

Madura is a regional language used by the people of the Madura Islands in Indonesia[1]. At present the use of Madura Language has begun to decrease. The existence of migrants from outside Java and the frequent communication in the community using the national language, Indonesian, is one of the causes. Communication in Madurese has a difference between children and adult[2]. There are 3 levels in Madura, namely: enja'-iyyah, engghi-enten, engghi-bhunten. The enja'-iyyah level is used for communication between friends or between people of the same age. The level of engghi-enten becomes the language of communication between people whose age differences are not too much different. While the engghi-bhunten level is used to communicate with people who are respected, who are much older, or who are considered elders.

In its pronunciation, Madura language has its own uniqueness like phonemes whose pronunciation is exhaled, namely phonemes: ‘bh’, ‘dh’, ‘gh’, dan ‘jh’[3]. In addition there were other problems in the Spelling of Madura Language [4]. The spoken vocals also have more than one spelling, such as: e with ɛ, a with ā. There is the quotation character (’), which is said at the end of the word, which means the...
The previous research has a translation system for others language. The translation system for English into Arabic has been done in 2005 [6]. Hindi language to English translation has also been carried out using a computer-based system [7]. Both of these systems produce translated text well. Nevertheless the output is only in the form of text. Text to speech system can be used to make it easy to convert text languages into voice pronunciation [8]. Text to speech system in Hindi language is done in 2014 [9]. An analysis of translator of Madura into English has been done by Masduki [10], but without a computerize system. With the spelling of Madurese language problem, a more complex system is needed. Online Indonesian-madurese text translation is also rarely found. Our research purpose text to speech application in the language translation system. Indonesian words or sentences can be translated in the form of text and sound. In this context, speech used as output is sound for each syllable of word. The method used for the translator is the Free Context Parsing Algorithm, while the chopping of words into syllables is done by applying the FSA. The organization of this paper is introduced in section 1. Section 2 is Madurese language. Section 3 is the Free Context Parsing Algorithm, while the chopping of words into syllables is done by applying the FSA. Section 4 is Madurese corpus expanding using speech of syllables. Section 5 is the Finite state automata method for Madurese word language. Section 6 is the Result and Discussion about the implementation of the system with three case studies. Section 7 is Conclusion. Others section are Reference and Acknowledgments.

2. Madurese Language

Madura language has special characteristics, both in the fields of phonology (sound language), morphology or syntax (arrangement of words or sentences). Its uniqueness includes: not knowing third-person pronouns, having inspired phonemes and tanaspirat, having morpheme functions “tang” or “sang”, and having special characters (e, ê, a, ū, bh, dh, gh, jh, quotation character (’)).

Examples of aspirate and tanaspirat phonemes:

- Tanaspirat: ɓūk (means ‘under’),
- Aspirat: ɓūb (means ‘onion’).

In the level of ‘enja’-iyyah’, the terms “tang” or “sang” are used as possessive pronouns, for example: “tang buku” not “b sën ‘”, “a eỹ së êkĭ êti sën ‘” (means this is that I like). The characters e with ê, and a with ū also differ in spelling. Spelling character is:

- a = read as usual as the word ‘awas’ in Indonesian
- ū = read as usual as the word ‘belajar’ in Indonesian
- e = read as usual as the word ‘kertas’ in Indonesia
- ê = read as usual as the word ‘bebas’ in Indonesia

Special characters bh, dh, gh, jh are read like bold, examples: ɓh ūjũ, dhũh, ghũhũy, hũjũhũ. Whereas the quoted characters read like the consonant ‘k’ but are read from the throat.

The formation of words in Madura is grouped into four types [11] namely: prefix, infix, suffix and conflation (prefix and suffix). The limitations on this research are limited to general words, not in abbreviated notes, absorption language, writing numbers or words that having an infix.

3. Proposed Method

This translation system is translated with a text to speech system. The input system that consisting of words or sentences are processed with preprocessing text, then next step is translation process. In this step, system needs Madura dictionary and Free Context Parsing algorithm. The preprocessing text are tokenizing and
stemming process. We use Enhanced Confix Stripping Stemmer (ECS) to stemming algorithm. The results of the translation process of Madurese text are used in the FSA algorithm to search for syllables from madurese words. The next process is text to speech process. The system looks for sound files in the corpus according to the syllables produced in the FSA process. The result of this system is the utterance of words or sentences that are synchronized with the text input. The translation scheme and the text to speech system are shown in Figure 1.

4. Expand of Madurese Corpus
The Madurese Dictionary is used in the process of translating Indonesian into Madurese. We use dictionary [3] which has translation words from Indonesian to Madurese in three variations of spelling variations. Three spelling variation of Madurese is: Enja’-iyyah, Engghi-enten, and Engghi-Bhunten. In the text to speech system, we need a corpus of Madurese syllables. In madurese dictionary we can not find this, so we need expand of madurese corpus with speech of syllable. This data was obtained with the help of Madurese people. The author records the sounds of possible syllable variations. This data was created manually using handphone, and successfully collected 1000 Madurese syllable speeches. The speech file is stored in the corpus in the form of a wav file.

5. Finite State Automata for Madurese
At this section, the FSA algorithm that used is explained. Section 5.1 explains the process of decapitating syllables of Madurese language. Section 5.2 explains the application of the FSA to the beheading of the syllable of Madurese.

5.1. Decapitation Process of Madurese syllables
Madurese language known seven (7) vowels (V), namely: a, å, e, ē, i, o, u and twenty five (25) consonants, namely: b, bh, c, d, dh, f, g, gh, h, j, jh, k, l, m, n, ng, ny, p, r, s, t, v, w, y, z. Consonants (C) bh, dh, gh, jh, ng, and ny are known as a consonantal letter. Quoted character (’) or Glotal (G) character is recognized as the letter itself.

Based on the rules of decapitating Indonesian [11], the rules for decapitating syllables in Madurese are explained as follows:

1. If at the beginning or in the middle of a word there are two vowels (V), then the decapitation is done between the two vowels (_V / V_)
2. If in the middle of the word there is a consonant letter (C) and a consonant combination between the two vowels, then the decapitation before the consonantal letter (_CV / CV_)

---

Figure 1. Scheme of Translation Indonesian-Madurese System
3. If in the middle of a word there are two or more consonantal letters, then the decapitation is done between the first consonant, including the consonantal letter combination, with the second consonant (\_C / CV\_).

4. If in the middle of a word there is a glotal (') (\_G\_) character, then the decapitation is done after the letter.

In accordance with the rules for decapitating of Madurese language, we make general patterns for decapitating Madurese language. This pattern will later be used to form states in the FSA algorithm. The pattern are:

1. V,
   example: a-têh (means: heart), a-ju-nan (means: you), ê-bhuh (means: mother)

2. VC
   example: ar-lo-jhi (means: wristwatch), an-dhuk (means: towel)

3. CV
   example: ka-lam-bhi (means: cloth), pa-lê-man (means: go home), sa-ê (means: delicious)

4. CVC
   example: ram-bhut (means: hair), bhun-ten (means: no), sep-po (means: old)

5. CCV
   Example: ghu-bhûr (means: go home), jhu-ghen (means: also)

6. VCC
   Example: êng-ghi (means: yes), ong-gha (means: up to)

7. CCVC
   Example: pot-trah (means: son), dhur-mas (means: shower)

8. CVCC
   Example: song-kan (means: sick), nang-kah (means: jackfruit)

9. CCVCC
   Example: ma-nyang (means: bee), am-bhung (means: nose)

10. CVG
    Example: o-ro' (means: massage), a-na' (means: son/daughter)

11. CCVG
    Example: dha'-'ar (means: eating), pêr-dhûr (means: short)

5.2. Finite State Automata for Syllables of Madurese

FSA in previous research [12] was used for decapitating Indonesian text syllables. Finite satate approach for spoken english is done [13]. FSA for machine translation can also be done [14]. In this research, FSA is used according to the general pattern in section 5.1 for decapitation of Madurese words. There are two interrelated FSA levels. The first level of FSA is recognize patterns V, G, C, CV, and CCV. The results of beheading at the first level become input for the second level FSA. In the second level FSA, system can recognize all the Madurese language patterns, which are 11 patterns. The FSA diagram in first level is shown in Figure 2. The second level of FSA diagram is seen in Figure 3.

6. Result and Discussion

The system test was carried out using three (3) case studies. The first case study system was tested for word beheading accuracy. The second case study was tested for the process of text to speech in Madura. From the translation process and the FSA, there are some words that can be beheaded at the first level, some are resolved at the second level, some must pass the first level and the second level. The third case is the trial of translation and text to speech in Madura.

Accuracy formula used is the calculation of the correct value of the total amount of data:

\[
A = \frac{s_t}{t} \times 100\%
\]  

(1)
Figure 2. FSA for syllables of Madurese in first level

Figure 3. FSA for syllables of Madurese in second level
Of the 50 madura test words in the first case found an accuracy value of 90%. The second case study of 20 words with 3 variations of spelling or a total of 60 data, it is known that the accuracy obtained is 100%. Case 3 of 20 Indonesian sentences with 3 variations of spelling found accuracy at 100% correct sound output but for intonation patterns the accuracy obtained was 85%. The value of accuracy for each case is shown in Table 1.

| Case | Total data | Accuracy value |
|------|------------|----------------|
| Case 1 | 50 words | 90% |
| Case 2 | 60 words | 100% |
| Case 3 | 20 sentences | 85% |

Figure 4 is an example of a test result for the word ‘andhuk’. This decapitation requires two levels of FSA. The results from the first level FSA are used as input for the second level FSA. In the first level FSA, the result of word beheading is: ‘a-n-duh-k’ while in the second level FSA is generated by the word beheading: an-dhuk.

![FSA Diagram](image)

**Figure 4.** FSA diagram for the word ‘andhuk’ from first level of FSA to second level of FSA

### 7. Conclusion

With this text to speech-based translation system using FSA, words or sentences in Indonesian can be translated into Madurese with various levels of language. From the test results, it can be seen that the FSA with the general pattern of Madurese words was found capable of converting words into syllables with an accuracy of 90%. For pronunciation of words with 3 levels are 100% correct. The problem is that the Spelling of madness is still 85%. This is possible because of the lack of corpus audio syllables of madurese. The limitation in this system is that the tests are limited to the data in the database and audio syllables of Madurese corpus and cannot be used for words based on absorption or the presence of infixes. Future research can be expanded towards expanding corpus of syllables of Madurese and developing modules for absorption words and infixes.
Acknowledgments
This research is one of the major research topics using Madurese as the data. The results of this study have been used and implemented on the website http://tts.madura.web.id. Some other research results related to the development of Madura Language applications are also on the website.

Reference
[1] M. Misnadin and J. Kirby, “Madurese,” *Journal of the International Phonetic Association*, pp. 1–18, 2018.
[2] I. Journal and H. Studies, “Madura language Variations in Children and Adults in Manduro Village seen from Phonological Change and Its Relationship with Language Change and Ethnic,” *International Journal of Humanity Studies*, vol. 2, no. 2, pp. 167–178, 2019.
[3] A. Prawitra, *Kamus Lengkap Bahasa Madura*. Jakarta: Dian Rakyat, 2009.
[4] A. Badrudin and A. Sofyan, “Problematics of Technical Preparation and Constraints in the Spelling System of Madurese Language,” *Parole: Journal of Linguistics and Education*, vol. 8, no. 2, pp. 52–56, 2018.
[5] W. D. Davies, *A Grammar of Madurese*. Berlin: Walter de Gruyter, 2010.
[6] A. T. Al-Taani and E. M. Hailat, “A Direct English-Arabic Machine Translation System,” *Information Technology Journal*, vol. 4, no. 3, pp. 256–261, 2005.
[7] E. Gupta and S. K. Shrivastava, “A Result Analysis of Translation Techniques of English to Hindi Online Translation Systems,” *International Journal of Computer Applications*, vol. 156, no. 12, pp. 12–15, 2016.
[8] I. Isewon, “Design and Implementation of Text To Speech Conversion for Visually Impaired People,” *International Journal of Applied Information Systems*, vol. 7, no. 2, pp. 25–30, 2014.
[9] K. Kamble and R. Kagalkar, “A Review: Translation of Text to Speech Conversion for Hindi Language,” *International Journal of Science and Research (IJSR)*, vol. 3, no. 11, pp. 1027–1031, 2014.
[10] J. Sastra, I. Fisib, and U. Trunojoyo, “The Translation of Madurese Cultural Expressions English,” *Jurnal Bahasa dan Seni*, vol. 44, no. 2, pp. 162–174, 2016.
[11] J. Parmin, “Menulis Ilmiah: Buku ajar MPK Bahasa Indonesia,” in *Buku ajar MPK Bahasa Indonesia*, Surabaya: Unesa University Press, 2011.
[12] E. S. Astuti, A. Naba, and W. A. Prijono, “Sistem TTS Dalam Bahasa Indonesia menggunakan Metode FSA dan Database Diphone,” in *Seminar Nasional Electrical, Informatics, and It’s Education*, 2009, pp. 139–144.
[13] H. Li, “A Novel Machine Translation Method Based on Stochastic Finite Automata Model for Spoken English,” *International Journal of Emerging Technologies in Learning*, vol. 14, no. 6, pp. 98–109, 2019.
[14] S. Bangalore, “A Finite-State Approach to Machine Translation,” 2002.