Robust Arabic Multi-stream Speech Recognition System in Noisy Environment

Anissa Imen Amrous and Mohamed Debyeche

Speech Communication and Signal Processing Laboratory (LPCTS),
Faculty of Electronics and Computer Sciences, USTHB
P.O. Box 32, Bab Ezzouar, Algiers, Algeria
amrous_im@hotmail.fr, mdebyeche@gmail.com

Abstract. In this paper, the framework of multi-stream combination has been explored to improve the noise robustness of automatic speech recognition systems. The main important issues of multi-stream systems are which features representation to combine and what importance (weights) be given to each one. Two stream features have been investigated, namely the MFCC features and a set of complementary features which consists of pitch frequency, energy and the first three formants. Empiric optimum weights are fixed for each stream. The multi-stream vectors are modeled by Hidden Markov Models (HMMs) with Gaussian Mixture Models (GMMs) state distributions. Our ASR is implemented using HTK toolkit and ARADIGIT corpus which is data base of Arabic spoken words. The obtained results show that for highly noisy speech, the proposed multi-stream vectors leads to a significant improvement in recognition accuracy.

Keywords: Multi-stream speech recognition, HMM, noisy environments.

1 Introduction

Improve the robustness of automatic speech recognition in presence of additive noise has become an active topic and a number of techniques has been proposed to improve word accuracies in noisy environments. The use of multi-stream models is one such technique [1]. A multi-stream speech recognizer is based on the combination of multiple feature streams each containing complementary information. The performance of such system depends on the fact that the selected features for every stream must not go through the same distortion in presence of noise. The weight given to each stream is another important aspect in multi-stream combination system. The rule should be such that the streams that are reliable should get more weight compared to the stream corrupted by noise [2], [3], [4].

We can refer to many works that tried to improve the robustness of ASR system by using several streams of features that rely on different underlying assumptions and exhibit different properties. Shimmer and jitter are used in [5], and formant and auditory-based acoustic cues are used together with MFCC in [6], [7]. In [8], [9], a multi-stream approach is used to combine MFCC features with formant estimates and

A. Elmoataz et al. (Eds.): ICISP 2012, LNCS 7340, pp. 571–578, 2012.

a selection of acoustic cues such as acute/grave, open/close, tense/lax, etc. Pitch has been also taken into account in many works for the recognition of tonal languages [10], [11]. For the same purpose, many works in audio-visual domain have investigated the contribution of the visual information on the acoustic recognition system in noisy environments [12], [13].

This work aims to improve ASR system in noisy environments by using a new multi stream vector based on MFCC, pitch, energy and the three first formants. The remainder of the paper is organized as follows: the multi-stream HMM based ASR systems are presented in section 2. In section 3, the experiments setup and results are given. Finally, we draw conclusions in Section 4.

2 Multi-stream HMM Based ASR System

The schematic overview of the multi-stream system is shown in *Fig.1*. Where γ_i (*i=1,2,..N*) is the stream weight of stream *i,* and it can be fixed statically [14] or estimated dynamically [3], [4]. Each stream is composed of a set of features and the *N* streams are combined to form a multi-stream vector at the input of the multi-stream modeling unit. In the test step the multi-stream features are decoded by a usual decoding ASR system unit.

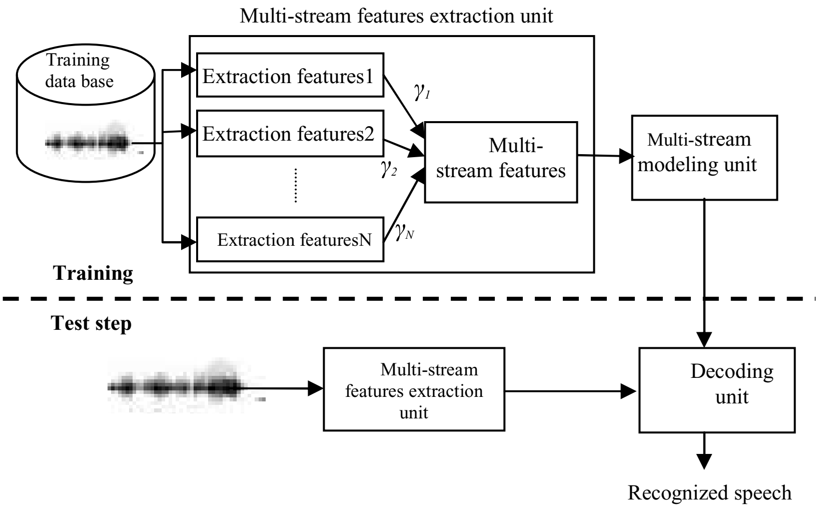

Fig. 1. Multi-stream HMM based ASR system

2.1 Multi stream Features

We describe in this section some of the theoretical background of the two stream features used in this work.

2.1.1 Streaml: MFCC Features

Our first stream is make up from the Mel-Frequency Cepstral Coefficients (MFCCs) [15] and their first (Δ) and second ($\Delta\,\Delta$) derivatives. For each analysis window, the MFCCs coefficients are calculated by equation (1), as follows:

$$MFCC(n) = \sum_{m=0}^{M-1} E[m] \cos\left(\frac{\pi n(m+\frac{1}{2})}{M} \right) \quad 0 \le n \le M \tag{1}$$

where M is the number of filter bank channels and *E[m]* is the energy of a given filter.

2.1.2 Stream 2: Complementary Features

The second stream consists of three kinds of features, namely pitch, energy and the first three formants. To complete the stream, the first and the second order derivatives of the five features are added.

According to the literature [16], [17], [18], those features are less affected by noise comparing to the usual features such as MFCC [15], PLP[19] and LPC [20] which represent the vocal tract characteristics and are very susceptible to noise.

2.1.2.1 Pitch. Its estimation is based on autocorrelation function [21].Giving a speech window $\{s(n), n = 0, 1, ..., N_{s-1}\}$ the autocorrelation function is defined as

$$R(k) = \frac{1}{N} \sum_{n=0}^{N_s = 1-K} s(n)s(n+k), \quad k = 0, ..., N_s - 1 \tag{2}$$

where N_s is the number of autocorrelation points to be computed.

2.1.2.2 Formant frequencies. In this paper we choose to use the frequencies of the first three formants which are estimated from the maxima of the LPC spectrum model [22]. These maxima are defined as the complex roots of the following polynomial:

$$1 + \sum_{i=1}^{P} a_i z^{-i} = 0 \tag{3}$$

where p is the LPC order.

2.1.2.3 Energy. Is defined as the variation of the signal amplitude caused by the force coming from the pharynx. The energy was computed by taking the logarithm of the windowed signal $(s_t)_{t=1,T}$ [23]:

$$E = \frac{1}{T}\sum_{t=1}^{T} s_t^2 \tag{4}$$

where T is the window signal $(s_t)_{t=1,T}$ size.

2.2 Multi-stream Modeling

A multi-stream model is a product model of the different feature streams. For *S* independent streams, the output distribution for state *j* using a Gaussian mixture is defined as

$$b_j(o_t) = \prod_{s=1}^{S}\left[\sum_{m=1}^{M_s} c_{jsm}\mathrm{N}(o_{st};\mu_{jsm},\sum\nolimits_{jsm})\right]^{\gamma_s} \tag{5}$$

where M_s is the number of mixture components for stream *s*, *cjsm* is the weight of the *m*-th component and N(o; μ, Σ) is a multivariate Gaussian with mean μ and covariance Σ [23]. The exponent γ_s is the weight for stream *s*.

2.3 Decoding

The decoding unit calculates the likelihood between the word to recognize and all the acoustic models which are already trained in the training step. The recognized word is the one which corresponds to the acoustic model according to the maximum likelihood. This likelihood was performed using the Viterbi algorithm [23].

3 Experimental Setup

This section presents the database and the experimental setup used for the evaluation of the proposal multi-stream HMM based ASR system.

3.1 Database Description

The speech database used in this work is the isolated ARADIGIT corpus [24]. It is composed of Arabic isolated digits from 0 until 9. This database is divided into the following corpuses:

- Train corpus: consisting of 1800 utterances pronounced by 60 speakers including the two genders, where, each speaker repeats the same digit 3 times.
- Test corpus: consisting of 1000 utterances pronounced by 50 speakers including the two genders, where, each speaker repeats the same digit 2 times.

3.2 Muti-stream Feature Extraction

- Stream1: For the first stream, MFCC features are extracted by HTK [23]. The speech signal is divided into a number of overlapping time windows of 25 ms with a frame period of 10 ms. For each analysis window, 12 MFCC features with their delta and acceleration coefficients, resulting in a feature vector of 36 acoustic features (MFCC_D_A) has been extracted.
- Stream2: The complementary features of the second stream which are: pitch, energy and the first three formants are extracted by the Praat package [25] based o n algorithms described in section 2.1.2. Delta and accelerations coefficients are added to this stream by HTK, making a total vector stream size of 15 coefficients (Comp_D_A).

3.3 Experimental Methodology

Our experiments were developed using HTK package (Hidden Markov Toolkit) [23], from Cambridge University. With the aim to show the advantage of using multi-stream features in speech recognition under real-life test conditions, we carried out a set of experiments. Four ASR systems are built:

1. **Single stream1 ASR system**: uses as observation vectors, features of stream1.
2. **Single stream2 ASR system**: uses as observation vectors, features of stream1.
3. **Equally-weighted multi-stream ASR system**: uses as observation vectors, features of stream1 concatenated to stream2 features. The two stream are equally-weighted ($\gamma_1 = \gamma_2 = 1$) .
4. **Optimally-weighted multi-stream ASR system:** uses as observation vectors, features of stream1 concatenated to stream2 features. The two streams are optimally weighted. The optimum weights are chosen empirically from experiences. Stream1 weights for each of the SNR's are as shown in Table 1. The weights of the second stream may be computed from this table using $\gamma_2 = 2 - \gamma_1$.

Table 1. Stream1 weights

SNR(dB)	20dB	15dB	10dB	5dB	0dB	-5dB
γ_1	1.1	1.1	1.1	1.1	0.9	0.9

The HMM models used for the all systems are a left-to-right HMM with continuous observation densities. Each model consists of 3 states, in which, each state is modeled by 1 Gaussian mixture with a diagonal covariance matrices defined as in equation (5).

To simulate the adverse conditions of test, we have corrupted the database by an airport noise extracted from the NOISEX92 database [26] and added to the speech signal with SNR ranging from -5 dB to 20 dB.

The acoustic models' training uses the clean speech database; the noise is only added for testing the recognition performance.

Table 2. Comparative speech recognition results

SNR (dB)	20dB	15dB	10dB	5dB	0dB	-5dB
Single stream1 ASR system	80.81%	67.99%	49.91%	30.35%	17.99%	9.96%
Single stream2 ASR system	59.32%	56.09%	52.82%	41.97%	26.66%	18.36%
Equally-weighted multi-stream ASR system	84.96%	75.92%	65.41%	46.68%	32.75%	20.76%
Optimally-weighted multi-stream ASR system	**85.89%**	**76.38%**	**66.05%**	**47.51%**	**32.93%**	**21.49%**

3.4 Results

Table 2 gives the results for the implemented ASR systems in different test conditions. Best results in terms of word recognition accuracy are edited in bold. For single stream systems, the ASR system based on stream1 (MFCC) outperform the one based on stream2 in quite noisy environments (20dB, 15 dB). In highly noisy environments, it is the single stream2 system which performs better than the single stream1 one. for instance, at 5 dB 41.97vs. 30.35. This is due to the fact that the proposed complementary features were more robust to noise comparing to the MFCC features.

As it can be observed, overall (SNR = -5 to 20 dB), the multi-stream systems, either equally-weighted or optimally-weighted, shows an improvements in word accuracy over the single stream systems. Another interesting aspect of these results is that the improvement in word accuracies is more pronounced in cases of low SNRs. For example, with 5dB : 47.51%% vs. 30.35%, i.e., an improvement of 17.16%is noticed.

It can be seen that the optimally-weighted system gives a better word accuracy when compared to the equally weighted system by about 1%. This shows the important role of weights in multi-stream framework.

4 Conclusions

In this paper, we have studied the contribution of a new multi-stream vector for Arabic speech recognition system based on Hidden Markov Model. The new multi-stream vector is consisted of the standard cepstral features MFCC, and a set of complementary features namely, pitch, energy and the first three formants. Results show that with these complementary features we can get significant word accuracy improvement over both single and multi stream ASR systems.

References

1. Janin, A., Ellis, D., Morgan, N.: Multi-stream speech recognition: ready for prime time. In: Proc. of Eurospeech, Budapest (1999)
2. Guo, H., Chen, Q., Huang, D., Zhao, X.: A Multi-stream Speech Recognition System Based on The Estimation of Stream Weights. In: Proc. ICISP, pp. 3479 – 3482 (2010)
3. Sanchez-soto, E., Potaminos, A., Daoudi, K.: Unsupervised stream weights computation in classification and recognition Tasks. IEEE Trans. Audio, Speech and Language Processing 17(3), 436–445 (2009)
4. Potamianos, A., Sánchez-Soto, E., Daoudi, K.: Stream weight computation for multi-stream classifiers. In: Proc. ICASSP, pp. 353–356 (2006)
5. Li, X., Tao, J., Johanson, M.T., Soltis, Savage, J.: Stress and emotion classification using jitter and shimmer features. In: Proc. ICASSP, vol. 4, pp. IV-1081–IV-1084(2007)
6. Holmes, J.N., Holmes, W.J.: Using formant frequencies in speech recognition. In: Proc. Eurospeech, Rhodes, pp. 2083–2086 (1997)
7. Selouani, S.A., Tolba, H.: Distinctive features, formants and cepstral coefficients to improve automatic speech recognition. In: Proc. IASTED, pp. 530–535 (2002)
8. Selouani, S.A., Tolba, H., O'Shaughnessy, D.: Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm. In: Proc. of ICASSP, pp. 837–840 (2002)
9. Tolba, H., Selouani, S.A., O'Shaughnessy, D.: Comparative experiments to evaluate the use of auditory-based acoustic distinctive features and formant cues for robust automatic speech recognition in low snr car environments. In: Proc. of Eurospeech, pp. 3085–3088 (2003)
10. Chongjia, N.I., Wenju, L., Xu, B.: Improved Large Vocabulary Mandarin Speech Recognition Using Prosodic and Lexical Information in Maximum Entropy Framework. In: Proc. CCPR 2009, pp. 1–4 (2009)
11. Ma, B., Zhu, D., Tong, R.: Chinese Dialect Identification Using Tone Features Based on Pitch Flux. In: Proc. ICASSP, p. I (2006)
12. Gurbuz, S., Tufekci, Z., Patterson, E., Gowdy, John, N.: Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition. In: Proc. ICASSP, pp. II-2021–II-2024 (2002)
13. Guoyun, L.V., Dongmei, J., Rongchun, Z., Yunshu, H.: Multi-stream Asynchrony Modeling for Audio-Visual Speech Recognition. In: Proc. ISM, pp. 37–44 (2007)
14. Addou, D., Selouani, S.A., Boudraa, M., Boudraa, B.: Transform-based multi-feature optimization for robust distributed speech recognition. In: Proc. GCC, pp. 505– 508 (2011)
15. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Proc. IEEE Trans. ASSP 28, 357–366 (1980)
16. Mary, L., Yegnanarayana, B.: Extraction and representation of prosodic features for language and speaker recognition. Proc. Speech Communication 50, 782–796 (2008)
17. Doss, M.: Using auxiliary sources of knowledge for automatic speech recognition. Ph.D Theses; École Polytechnique Fédérale de Lausane (2005)

18. Ververidis, D., Kotropoulos, C.: Emotional speech recognition: resources, features, and methods. Proc. Speech Communication 48, 1162–1181 (2006)
19. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America 87, 1738–1752 (1990)
20. Slifka, J., Anderson, T.R.: Speaker modification with lpc pole analysis. In: Proc. of ICASSP, pp. 644–647 (1995)
21. Rabiner, L.R.: On the Use of Autocorrelation Analysis for Pitch Detection. IEEE Transaction on Acoustics, Speech, and Signal Processing 25, 1 (1977)
22. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences. IEEE Trans. on Speech and Audio Processing 28(4), 357–366 (1980)
23. Young, S., Odell, J., et al.: The HTK Book Version 3.3. Speech group, Engineering Department. Cambridge University Press (2005)
24. Amrouche, A.: Reconnaissance automatique de la parole par les modèles connexionnistes. Ph.D Theses, Faculty of Electronics and Computer Sciences, USTHB (2007)
25. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (2008), http://www.praat.org/
26. Varga, A.P., Steeneken, H.J.M., et al.: The NOISEX-92 study on the effect of additive noise on automatic speech recognition. In: NOISEX 1992 CDROM (1992)